Title

The Effect of Form-Meaning Consistency on Word Learning Through Reading: Are Pseudo-Neighbors Harder to Learn?

Author

Wang-Kildegaard, Bowen

Publication Date

2024

Peer reviewed|Thesis/dissertation

The Effect of Form-Meaning Consistency on Word Learning Through Reading: Are Pseudo-Neighbors Harder to Learn?

By

Bowen Wang-Kildegaard

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anne Cunningham, Chair
Professor Sophia Rabe-Hesketh
Professor Mahesh Srinivasan

Spring 2024

Abstract

The Effect of Form-Meaning Consistency on Word Learning Through Reading: Are Pseudo-Neighbors Harder to Learn?

by

Bowen Wang-Kildegaard

Doctor of Philosophy in Education

University of California, Berkeley

Professor Anne Cunningham, Chair

Efficient language processing necessitates some level of systematic mapping between word forms and meanings (Kirby et al., 2008; Dautriche et al., 2017). When words look similar but have unrelated meanings (e.g., 'leopard' vs. 'leotard'), this inconsistent form-meaning mapping could interfere with word recognition (Forster & Hector, 2002; Bowers et al., 2005). For example, the activation of 'leopard' when seeing 'leotard' interferes with judging whether 'leotard' is an animal (Rodd, 2004). Little is known about whether similar interference occurs at the early stages of word learning.

In this study, I propose the form-meaning consistency (FMC) model, which offers a fine-grained analysis of the semantic relationships between words with similar forms. For example, 'leopard-leotard' exhibits low FMC due to similar forms but unrelated meanings, whereas 'snort-snore' shows high FMC with closely related forms and meanings.

This study examined the effect of form-meaning consistency on word learning (e.g., Does prior knowledge of 'leopard' interfere with learning 'leotard'?). I designed and conducted a novel experiment where 50 adult native English speakers read short passages to learn 12 pseudowords that differ in their form-meaning consistency with known words. For example, I investigated whether the activation of 'bucket' interfered with learning the meaning of 'burket,' a fictional animal, in the early stages of acquisition. Learning was measured by semantic relatedness rating tasks.

The findings indicate that lower FMC interfered with the semantic learning of novel words. Specifically, the activation of semantically less related orthographic neighbors led to longer response time and biased ratings in semantic tasks. This interference effect was most salient in delayed semantic tasks and was mitigated in immediate tasks. The analysis reveals various mechanisms through which FMC influences the retrieval of word meaning and decision on semantic relatedness.

This study contributes to the understanding of how the brain's semantic network processes and integrates novel linguistic information, highlighting the significance of form-meaning mappings

in efficient language learning and usage. By extending the current understanding of the role of form-meaning mapping in word recognition to the realm of word learning, this research provides insights into the mechanisms and cognitive processes underlying vocabulary learning from reading. It also has significant implications on more effective language teaching strategies, particularly in designing learning materials that minimize interference and optimize learning.

Keywords: Word learning; Form-to-Meaning Mapping; Vocabulary; Reading; Word Processing; Systematicity

DEDICATION

Dedicated to my husband, Christopher Wang-Kildegaard, who loves me unconditionally and taught me how to love unconditionally.

TABLE OF CONTENTS

Abstract 1
DEDICATION i
ACKNOWLEDGMENTS iv
CHAPTER I: INTRODUCTION 1
Background 1
Introducing the Form-Meaning Consistency (FMC) Model 3
Quantifying Form-Meaning Consistency 6
Theoretical Significance 7
Present Study 8
Research Questions 8
Hypotheses 8
CHAPTER II: METHOD 9
Experimental Design 9
Counterbalancing 11
Experimental Stimuli 13
Chimeras 13
Reading Materials 13
Probes 15
Pseudowords 15
Participants 17
Procedure 17
Task-Specific Hypotheses 19
Statistical Analyses 25
CHAPTER III: RESULTS 26
High FMC Against Different Probes 27
Ratings 30
Response Time 30
Medium FMC Against Different Probes 31
Ratings 33

Response Time 35
Low FMC Against Different Probes 36
Ratings 38
Response Time 39
Low Probe Against Different FMC Conditions 39
Ratings 41
Response Time 42
Medium Probe Against Different FMC Conditions 43
Ratings 45
Response Time 46
High Probe Against Different FMC Conditions 46
Ratings 48
Response Time 49
CHAPTER IV: DISCUSSION 50
The Effect of FMC on Ratings and Response Time 50
Mitigated Effect of FMC in the Immediate+Delayed Group 52
Additional Mechanisms That Influenced Ratings and Response Times 53
Addressing Alternative Hypothesis 55
Connection with Existing Literature 56
Conclusion 56
Limitations and Future Directions 57
Significance 58
REFERENCES 60
Supplemental Material: Model Diagnostics 64
Supplemental Material: Experimental Stimuli 70

ACKNOWLEDGMENTS

I am immensely grateful for the guidance and support of all my mentors throughout my academic journey.

Anne Cunningham: Since I began my PhD program in 2018, you have been an inspiring advisor. As a young adult who came to a foreign country to pursue doctoral studies in a second language, I was not confident in myself. Your positive feedback and encouragement not only gave me a lot of strength and confidence but also played a pivotal role in shaping my identity as a scholar. You inspired me to immerse myself in the field so that I could make contributions that are both innovative and rigorous. Your affirmation of my scholarly potential will continue to motivate me to pursue a rigorous program of research and strive for excellence in the years to come. Additionally, I deeply appreciate your invaluable guidance on my professional development, both within and beyond academia. You have provided me with a wide range of opportunities, such as collaborating with educational organizations to translate research into practice and engaging in peer review for journals. Your guidance always nudges me in the right direction, giving me the courage to apply for grants and submit my works to conferences and journals. All of these experiences have shaped who I am today, for which I am and will always be grateful.

Sophia Rabe-Hesketh: I am forever grateful for the opportunity to learn from you and to serve as a graduate student instructor for your statistics courses. Five years ago, when I was struggling to secure a position, you offered me the position in your Data 2 course, which meant a great deal to me. Your faith in me, despite my limited teaching experience back then, inspired me to work hard to be the best teacher possible and to honor your trust. Your passion and conscientiousness in teaching have been my role models. Having had the privilege of working with you for three years, I have gained significant pedagogical skills and statistical expertise, enabling me to successfully help my students thrive. Without your guidance and mentorship, I would not have acquired the ability to conduct the sophisticated statistical analyses presented in this dissertation. Inspired by your example, I am committed to helping more students master statistics and continuing to apply statistical methods in my future research.

Mahesh Srinivasan: You have been an essential mentor in my research journey. Thank you very much for welcoming me to your research group when I reached out four years ago. It has been a privilege to learn from you and collaborate with you on research. Your constructive feedback has continually challenged me to think critically and elevate my research to the highest academic standards. I look forward to continuing our collaborations on research projects that unite our interests and contribute to the field.

Last but not least, I want to express my gratitude to all my peers who shared this journey with me. We exchanged constructive feedback, shared resources, and provided emotional support to each other. I will always hold our friendship dear to my heart.

CHAPTER I: INTRODUCTION

Background

The association between the sound of a word and what that word refers to is mostly arbitrary in human languages (Saussure, 1916). This arbitrary sound-meaning association is evident in that different languages could refer to the same thing with radically different sounds. However, within each language, the association between sound and meaning is not entirely random (Louwerse & Qu, 2017). During the evolution of each language, some systematicity of how sound maps to meaning emerged (Kirby et al., 2008). Research has found that within a specific language, if a pair of words are semantically similar to each other, they also tend to be more phonologically similar; this pattern has been found in more than 100 languages including English (Dautriche et al., 2017). Research has found that nouns and verbs have distinct phonological features in English (Kelly, 1992; Farmer et al., 2006), which also supports the systematic mapping between sound and meaning (i.e., systematicity). Using this systematicity, computational simulations can categorize nonwords into nouns and verbs purely on the basis of the nonwords' sounds, and this categorization agrees with human judgments (Cassani et al., 2020).

These findings may be surprising to many because one can think of many examples where two English synonyms have very different sounds (e.g., 'big' vs. 'large') or examples where two phonological neighbors have very different meanings (e.g., 'hat' vs. 'hate'). However, what these studies found is an overall statistical trend across an entire language that outweighs specific counterexamples. Languages may have evolved in this way so as to reduce the processing load of language usage (Dautriche et al., 2017). Therefore, one may wonder whether word processing and learning will be less efficient in cases where sound and meaning are not consistently mapped.

Besides the *oral* form of words (sound/phonology), how consistently the *written* form (spelling/orthography) is mapped to meaning may also affect word processing and learning. In one's mental lexicon (storage of word knowledge in one's mind), words are represented in a network where they are connected with each other through meaning and form (Stella et al., 2018) instead of being stored in dictionary-like entries listing individual orthographic, phonological and semantic information. Semantics are involved early on in visual word recognition (Pecher, 2001). When one is trying to recognize a word, the orthographic and semantic representations of that word's orthographic neighbors are also activated (Rodd, 2004). For instance, during the process of recognizing 'leotard', the form and meaning of 'leopard' are activated simultaneously (Rodd, 2004); 'hat' is activated when recognizing 'hatch' and 'chat', and interferes with the semantic processing of 'hatch' and 'chat' (Bowers, et al., 2005)

When two words look alike, their meanings may or may not be related to each other. For instance, 'dealer' and 'deal' look alike and are morphologically related (i.e., adding an -er suffix to 'deal' makes 'dealer') so they are semantically related as well. In contrast, 'corner' and 'corn' look alike and are ostensibly morphologically related ('pseudo-morphologically related') but are not semantically related. Pairs such as 'scandal' and 'scan' look alike but are neither morphologically/psuedomorphologically nor semantically related. In masked priming lexical decision tasks where one needs to quickly decide whether a word is a real word, priming 'deal' with 'dealer' leads to faster recognition of 'deal' as a real word. The faster recognition of the target word caused by the presentation of a prime word is called the priming effect. Priming 'corn' with 'corner' also leads to faster recognition of 'corn' but the priming effect is not as strong. The priming

effect of priming ‘scan’ with ‘scandal’ is even less strong but is still present. This line of research has been conducted in different languages (e.g., Longtin et al., 2003; Rastle et al., 2004) and indicates that the activation of the meaning of a word occurs early on in word recognition and that the priming effect caused by orthographic neighbors is moderated by the degree to which the orthographic neighbor is semantically/morphologically related to the target word.

Marelli et al. (2015) put forth the concept of orthographic-semantic consistency (OSC) to explain this phenomenon. Letter strings like ‘widow’ have high OSC because most words containing the letter string ‘widow’ (e.g., widower, widowhood, widowed) are related to the concept WIDOW. In contrast, letter strings like ‘whisk’ have low OSC because most words containing the letter string ‘whisk’ (e.g., whisker, whiskered, whiskery) are not related to the concept WHISK. Therefore, the mapping between the letter string ‘widow’ and the meaning WIDOW is more consistent than the mapping between the letter string ‘whisk’ and the meaning WHISK (Marelli & Amenta, 2018). OSC quantifies how consistently a specific letter string maps to a specific meaning by calculating the semantic similarity between that letter string (e.g., ‘whisk’) and all the words containing that complete letter string (e.g., ‘whisker,’ ‘whiskered,’ etc.). They quantified semantic similarity using a distributional semantic model called the Continuous Bag of Words model. High OSC means that a letter string is a reliable cue for meaning, whereas low OSC means that a letter string is not a reliable cue for meaning. In other words, one can be relatively confident that whenever they see the high-OSC letter string ‘widow’ in a word, that word’s meaning should be related to WIDOW. In contrast, when one sees the low-OSC letter string ‘whisk’ in a word, the probability of that word referring to the concept of WHISK is low.

Marelli and Amenta (2018) hypothesized that OSC may predict word recognition response time for the following reasons. When recognizing a word, the orthographic and semantic representations of that word’s orthographic relatives are also activated. Orthographic relatives refer to words that contain the complete letter string of the target word. For a word with low OSC, because the meanings of its orthographic relatives are unrelated to it, the activation of the diverging semantic representations of these orthographic relatives may interfere with the recognition of the target word due to competition. In contrast, for a word with high OSC, the converging semantic representations of orthographic relatives with related meanings may facilitate the recognition of the target word. The effect of OSC is similar to the semantic ambiguity effect where word recognition is slower when a word has multiple unrelated meanings than when it has multiple related senses (Rodd et al., 2002).

Several recent studies found that OSC is indeed a significant predictor of reaction times in lexical decision and word naming tasks. Specifically, the higher the OSC of a word is, the faster the respondents can recognize that word in unprimed lexical decision tasks, after controlling for the word’s family size (i.e., the number of words that are morphologically related to the target word), length, and frequency (Marelli et al., 2015; Marelli & Amenta, 2018). Additionally, OSC correlates only weakly with these control variables. These results hold in both American English and British English (Marelli & Amenta, 2018), indicating that OSC captures a novel source of variance in visual word recognition.

Similarly, Amenta et al., (2017) developed a phonology-semantics consistency (PSC) measure as a phonological counterpart of OSC. PSC of a target word is the frequency-weighted average of the semantic similarities between that word and its ‘phonological relatives’ (for instance, ‘cognac’-/ˈkɒnjæk/ is a phonological relative for ‘yak’-/’jæk/). They have found that both OSC

and PSC explain variance in lexical decision time; there is an interaction between OSC and PSC where the effect of OSC is stronger when PSC is lower and vice versa. They have also found that the effect of OSC is largely mediated by PSC. In other words, the activation of the semantic representations of a word's orthographic relatives is largely through an orthography-phonology-semantics pathway; the direct orthography-semantics pathway is secondary but plays a larger role when the phonology-semantics mapping is less consistent. These findings suggest that consistent mapping between words' form and meaning contributes to efficient word recognition.

To sum up, existing research suggests that consistent mapping between words' form and meaning contributes to efficient word recognition. However, little is known about how consistency in form-meaning mapping may affect *learning* a new word. For example, does knowing the word 'whisker' interfere with learning the word 'whisk'?

Introducing the Form-Meaning Consistency (FMC) Model

Inconsistent form-meaning mapping may hinder word learning. When learning a new word through reading, if the new word contains a known letter string or looks/sounds like a known letter string but the meaning of the new word is *unrelated* to the known letter string, the learning of this new word's meaning may be interfered with by the activation of the semantic information of the known letter string. For instance, when one knows 'whisker' (facial hair) and is trying to learn the new word 'whisk' (mix), the activation of the meaning of 'whisker' may interfere with learning the meaning of 'whisk.' In contrast, if the meaning of the new word is *related* to the known letter string, the learning of the new word's meaning may be facilitated. For instance, the activation of the meaning of 'widow' may facilitate learning the meaning of 'widowed.' This interference or facilitative effect may not be limited to words sharing the same letter string but also occurs for words that are orthographically or phonologically similar to each other. For instance, 'leopard' is orthographically similar to 'leotard,' but the two words are semantically unrelated; thus, the semantic learning of 'leotard' may be interfered with by the activation of 'leopard.'

Existing research has found an inhibitory effect of semantically-unrelated orthographic neighbors on word recognition in semantic tasks among adult native speakers (e.g., Forster & Hector, 2022; Rodd, 2004; Bowers, et al., 2005). In Forster and Hector (2002), participants were asked to judge whether a set of real words and nonwords are animals, where the only correct response to nonwords was 'not animals.' They found that when the nonwords have a real-word orthographic neighbor that is an animal (e.g., turple), the participants took longer to reject these nonwords and made more errors than when the nonwords do not have an animal neighbor (e.g., tabric). This result indicates that the semantic representation of 'turtle' is activated when seeing 'turple,' thus delaying the rejection of 'turple.'

A similar result has been found for real words. In Rodd (2004), participants were asked to judge whether words were animal names (in Experiment 1A) or plant names (in Experiment 1B). The experimental items (e.g., leotard) are words whose only orthographic neighbor was an animal (e.g., leopard) whereas the control items (e.g., cellar) were words whose only orthographic neighbor was not an animal (e.g., collar). In Experiment 1A (i.e., animal categorization), responses to the experimental words (e.g., leotard) were 72 msec slower than those to the control words (e.g., cellar), on average. In Experiment 1B (i.e., plant categorization), average response times were 15 msec longer for the experimental words than for the control words. These results can be explained

by the interference caused by the activation of the animal neighbors of the experimental items. However, whether similar phenomena can be observed during the early stages of word learning has yet to be examined by empirical studies.

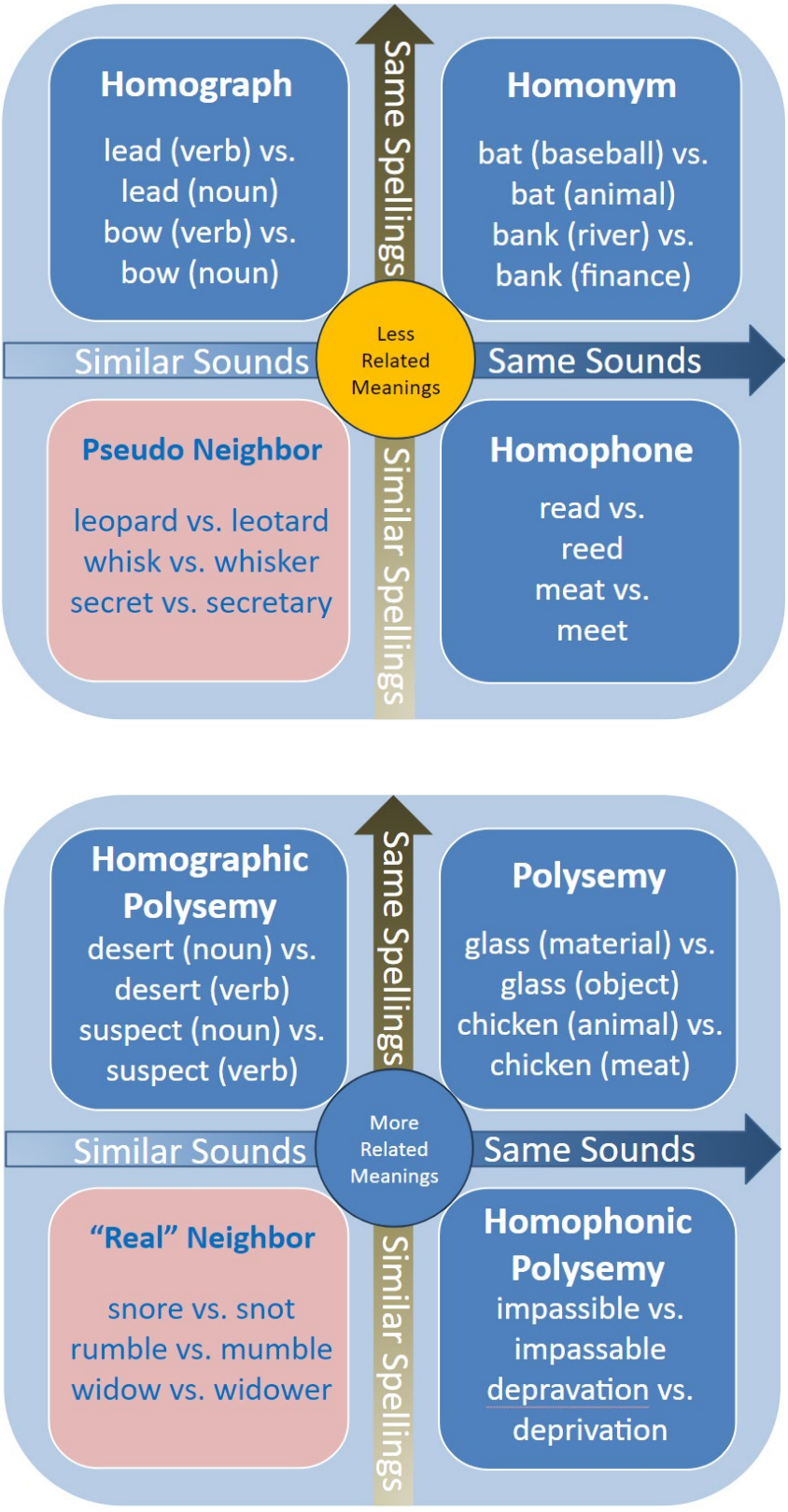
I have developed the Form-Meaning Consistency (FMC) model to test these hypotheses. This model categorizes word pairs into eight types based on their form (spelling and sound) and meaning relationships. Figure 1 demonstrates two broad categories: consistent and inconsistent form-meaning mappings. Inconsistent mappings have similar or identical forms but less related meanings, while consistent mappings have more related meanings. Inconsistent mappings include four types: **pseudo-neighbor** (similar spellings and sounds), homonym (identical spellings and sounds), homophone (identical sounds and similar spellings), and homograph (identical spellings and similar sounds). Consistent mappings also include four types: **'real' neighbor** (similar spellings and sounds), polysemy (identical spellings and sounds), homophonic polysemy (identical sounds and similar spellings), and homographic polysemy (identical spellings and similar sounds).

While traditional taxonomy only considers word pairs with *identical* forms, such as homonyms, the FMC model includes word pairs with *similar* forms. Most notably, the FMC model contributes two novel types: **pseudo-neighbors** and **'real' neighbors**. Examples of pseudo-neighbors include 'whisk-whisker' and 'leopard-leotard'; examples of 'real' neighbors include 'widow-widower' and 'snort-snore.'

Note that the pairwise form-meaning consistency framework here is slightly different from how orthographic-semantic consistency (OSC) is operationalized in Marelli & Amenta (2018). FMC is the semantic similarity between a pair of two words that have the same or similar forms, whereas OSC is the average semantic similarity between a target word and all other words that contain that target word. That being said, the underlying construct is arguably the same, which is the extent to which a specific word form is a reliable cue for a specific meaning. OSC is suitable for studying adults' word recognition because skilled adult readers would have learned all the words they are supposed to know. Therefore, it makes sense to consider all the orthographic relatives of the target word. However, if we are interested in studying how the knowledge of a new word is influenced by the existing knowledge of a known word, the pairwise comparison may be more relevant.

Figure 1

Taxonomy of Form-Meaning Consistency and Inconsistency



Note. Each figure has four quadrants, based on whether the spellings and sounds of the word pair are similar or the same. The top figure categorizes word pairs with less related meanings (lower form-meaning consistency) whereas the bottom figure categorizes word pairs with more related meanings (higher form-meaning consistency). The red box in the top figure represents "pseudo neighbors" with similar sounds and spelling but less related meanings. The red box in the bottom figure represents "real neighbor" with similar sounds and spellings as well as more related meanings.

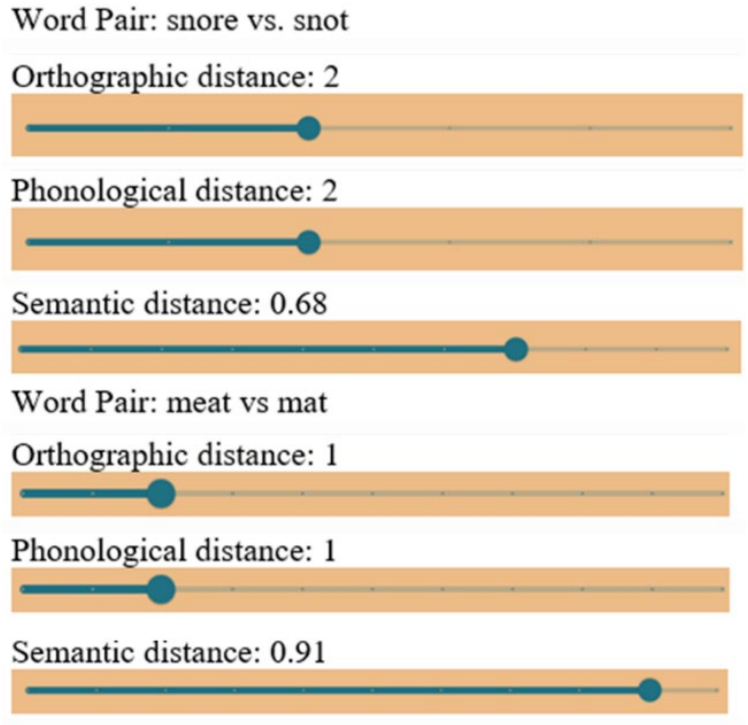
Quantifying Form-Meaning Consistency

In contrast to existing taxonomies that often depend on subjective criteria, the FMC model incorporates quantitative measures for orthographic, phonological, and semantic similarities. Orthographic distance is calculated using the Levenshtein distance metric, the least number of single-letter changes needed to change one word into another. Phonological distance utilizes a phoneme-based variant of the Levenshtein distance, which is operationalized as the least number of single-phoneme changes needed to change one word into another.

For semantic distance, I employ the distributional Semantic Models. These models are computational implementations of the *distributional hypothesis*, which argues that semantically similar words tend to appear in similar contexts (Harris, 1954). For instance, "teacher" and "instructor" have similar meanings because they appear frequently in similar contexts containing words like "school," "class," and "student". These models derive semantic representations as numeric vectors by abstracting from the distributional patterns of words across many contexts. Prominent examples include the Latent Semantic Analysis model (Landauer & Dumais, 1997) and the word2vec model (Mandera et al., 2017). A growing body of psychological research has shown that these models perform well in predicting behavioral data, supporting their psychological plausibility (Baroni et al., 2014; Günther et al., 2016; Mandera et al., 2017). Figure 2 offers illustrative examples, using the word2vec model to quantify semantic distances on a scale from 0 to 1, where a smaller number indicates greater similarity or relatedness.

Figure 2

Quantifying Form-Meaning Consistency



Theoretical Significance

The FMC model integrates various research topics and provides a comprehensive framework to investigate how existing knowledge impacts new word learning. Previous studies have examined the learning of polysemy and homonymy. A polysemous word is a word with multiple related senses, such as 'chicken', which can refer to both the living animal and the edible meat. In contrast, a homonym is a word with multiple unrelated meanings, such as 'pen', which can refer to either the writing instrument or animal enclosure. Existing research found that knowing one sense of a polysemous word aids in learning its other related senses (Floyd & Goldberg, 2021; Srinivasan et al., 2017, 2019; Srinivasan & Rabagliati, 2021; Srinivasan & Snedeker, 2011). Conversely, knowing one meaning of a homonym/homophone/homograph can hinder learning its other, unrelated meanings (Casenhiser, 2005; Fang et al., 2017; Mazzocco, 1997; Rodd et al., 2012; Saemen, 1970). In these cases, at least one of the written forms and the oral forms of the target word pair are the same. In other words, existing studies have examined, without referring to the construct of form-meaning consistency, how form-meaning consistency affects the learning of word pairs that have the *same* forms. However, the current literature has yet to explore the impact of form-meaning consistency on learning words with *similar* forms. Questions such as "Does knowledge of 'leopard' impede learning 'leotard'?" or "Does knowing 'snort' facilitate learning 'snore'?" remain unexplored.

Present Study

Research Questions

Building on the gaps identified in existing research and models, this study addresses the following research question: How does form-meaning consistency (FMC) between a novel word and a known word affect the learning of the novel word's meaning in the initial stages of learning?

Hypotheses

I hypothesize that if two words are orthographically and phonologically similar to each other but semantically distant, the semantic learning of the new word may be initially inhibited, similar to the 'clinging to primary meaning' effect when learning homonyms. For instance, if a person knows 'leopard' but not 'leotard' and encounters 'leotard' for the first time in their reading of a story, the semantic information of 'leopard' will be activated, and they may be biased towards associating the two when trying to infer the meaning of 'leotard.' The interference may even occur when the story context suggests otherwise. Even when learners can successfully infer the correct meaning of 'leotard' using contextual clues during reading, the interference may still be present when they try to retrieve the knowledge later. When they are subsequently tested on 'leotard,' 'leopard' will be activated and may interfere with recalling the meaning of 'leotard,' because the semantic representation of 'leotard' is not yet fully crystallized. On the other hand, existing knowledge of 'snore' may facilitate the inference and recall of 'snort.'

CHAPTER II: METHOD

Experimental Design

To empirically investigate my hypotheses on how FMC influences word learning, this study employed a behavioral experiment. Adult native English speakers learned the meanings of pseudowords by reading sentences embedded with these words.

Each pseudoword refers to a novel concept created by Lazaridou et al. (2017) where they combined pairs of related but distinct concepts (animals, plants, or objects) to form 'chimeras.' For instance, the 'alligator-rattlesnake' chimera refers to a fictional creature that has features of both alligator and rattlesnake. I adapted these chimeras and their reading materials from Lazaridou et al. (2017).

The originality of my study lies in my innovative experimental paradigm specifically designed to investigate FMC. I created pseudowords with varying levels of FMC to examine their impact on word-learning tasks. For example, for the 'alligator/rattlesnake' chimera, I have:

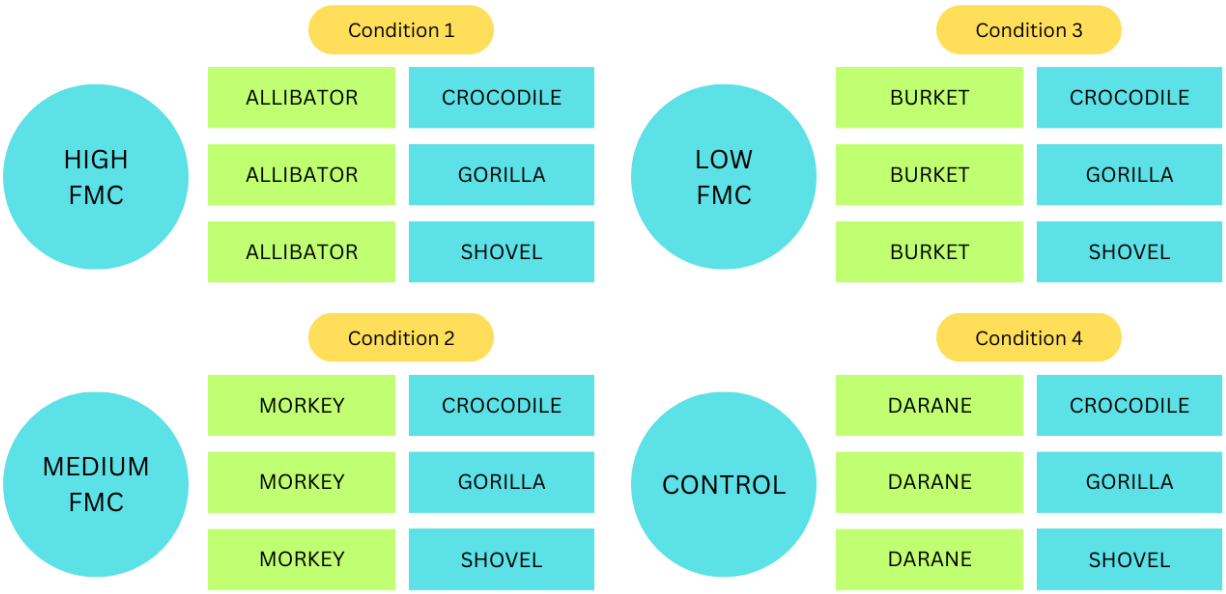
1. **High FMC**: 'allibator' (created from 'alligator,' whose meaning is highly related to the chimera and belong to the same broad semantic category, animal)
2. **Medium FMC**: 'morkey' (created from 'monkey,' whose meaning is moderately related to the chimera but still belong to the same broad semantic category, animal)
3. **Low FMC**: 'burket' (created from 'bucket,' whose meaning is the least related to the chimera and belongs to a different semantic category, object)
4. **Control**: 'darane' (randomly generated, with no real-word orthographic neighbor)

The participants engaged in semantic relatedness rating tasks post-reading. They rated the semantic relatedness between the pseudoword and three real words ('probe words') in random order. These probe words vary in semantic relatedness to the chimera and the pseudowords' base words. For instance, the probe words for the alligator/rattlesnake chimera include 'crocodile,' 'gorilla,' and 'shovel,' which closely relate to the base words 'alligator,' 'monkey,' and 'bucket,' respectively.

Figure 3 demonstrates the experimental design.

Figure 3

Experimental Design Diagram



Note. Under each condition, the green boxes represent the pseudoword for that condition whereas the blue boxes represent the probe words in the semantic relatedness rating tasks.

This design allows us to test the following hypotheses.

1. **High FMC may facilitate learning**: For instance, encountering 'allibator' may activate 'alligator,' which is semantically related to the alligator/rattlesnake chimera. This high consistency can facilitate the inference and retrieval of the correct meaning and speed up response time (see Task-Specific Hypotheses for detailed discussion of the hypothesized cognitive mechanisms behind this prediction).
2. **Medium FMC may interfere with learning moderately**: Encountering 'morkey' may activate 'monkey,' which is less semantically related to the chimera. This decreased consistency can lead to biased relatedness ratings and longer response times. For instance, the semantic relatedness between 'morkey' and 'gorilla' may be slightly but systematically overestimated, whereas the semantic relatedness between 'morkey' and 'crocodile' may be slightly but systematically underestimated.
3. **Low FMC may interfere with learning considerably**: Encountering 'burket' may activate 'bucket,' which is the least semantically related to the chimera. This low consistency can lead to the most bias in relatedness ratings and the longest response time. For instance, the semantic relatedness between 'burket' and 'shovel' may be overestimated whereas the relatedness between 'burket' and 'crocodile' and the relatedness between 'burket' and 'gorilla' may be underestimated.

More detailed task-by-condition hypotheses are under **Task-Specific Hypotheses**.

Counterbalancing

I used a Latin Square design to counterbalance the FMC conditions, as illustrated in Table 1. Each of the four participant groups was assigned a specific FMC condition for each chimera. For example, Participant Group 1 was assigned the control condition for the alligator/rattlesnake chimera, with the pseudoword ‘darane.’ Meanwhile, Participant Group 2 receives the high FMC condition for this chimera, with the pseudoword ‘allibator.’

Each group experienced all four FMC conditions but with different chimeras. In other words, I manipulated form-meaning consistency both within items across different participant groups and within participants across different items. Participants were pseudorandomly and evenly distributed among the four groups. One Latin Square was generated for each set of four chimeras (including four animals, four plants, and four objects, see Table 1). For each participant, the order of the 12 chimeras was randomized in the experiment.

Table 1

Counterbalancing FMC across Chimeras across Participants

	Alligator/ Rattlesnake	Elephant/ Bison	Peacock/ Goose	Caterpillar/ Cockroach
Group 1	Control (darane)	High FMC (elethant)	Medium FMC (cheepah)	Low FMC (cothage)
Group 2	High FMC (allibator)	Low FMC (wronch)	Control (gleadop)	Medium FMC (harster)
Group 3	Low FMC (burket)	Medium FMC (mostuito)	High FMC (pescock)	Control (teissem)
Group 4	Medium FMC (morkey)	Control (naisern)	Low FMC (garnic)	High FMC (coctroach)

	Car/ Van	Train/ Bus	Dishwasher/ Oven	Cannon/ Rifle
Group 1	Medium FMC (bicacle)	Low FMC (scortion)	Control (nefrim)	High FMC (rikle)
Group 2	Low FMC (capary)	High FMC (shultle)	Medium FMC (baurel)	Control (rordin)
Group 3	High FMC (caranon)	Control (nacrut)	Low FMC (japuar)	Medium FMC (dagrer)
Group 4	Control (thrafel)	Medium FMC (fefry)	High FMC (furtace)	Low FMC (coupar)

	Potato/ Turnip	Cucumber/ Celery	Corn/ Yam	Broccoli/ Spinach
Group 1	High FMC (porato)	Medium FMC (mungo)	Low FMC (zegra)	Control (segost)
Group 2	Medium FMC (oradge)	Control (gemack)	High FMC (pumpsin)	Low FMC (bontle)
Group 3	Control (vernag)	Low FMC (pirlow)	Medium FMC (chorry)	High FMC (spirach)
Group 4	Low FMC (sasmon)	High FMC (cekery)	Control (gitid)	Medium FMC (balana)

Experimental Stimuli

Chimeras

I adopted the chimeras devised by Lazaridou et al. (2017). They devised each chimera by matching a basic-level concrete concept, termed the *pivot*, with a semantically *compatible term*. The compatible terms were selected from the top 10 words with the highest semantic similarity to the pivot, based on McRae et al.'s (2005) word similarity norms. The pivot's synonyms, closely related co-hyponyms, and hyper/hyponyms were excluded from consideration. For example, 'alligator' pairs with 'rattlesnake' to form the 'alligator/snake' chimera. Twelve chimeras were used in the present study (see 'Probe' section for details on how chimeras were selected on basis of the quality of probes).

Reading Materials

The reading materials were adapted from Lazaridou et al. (2017). They generated 'passages' for each chimera, drawing an equal number of sentences from the British National Corpus and ukWaC, which are representative of written/spoken English and web texts, respectively. Each 'passage' comprises six sentences: three featuring the pivot word (e.g., 'alligator') and three with the compatible component (e.g., 'rattlesnake'). After replacing these terms with a pseudoword, the sentences collectively describe a unified entity. Averaging 17.6 words per sentence, the 'passages' offer reasonably informative clues about the referent type (e.g., a land-dwelling animal) without being so explicit as to reveal its exact identity.

For each chimera, I curated ten sentences (five for each chimera component) from Lazaridou et al.'s (2017) pool of 60 sentences, adhering to the following criteria:

1. They do not contain overt linguistic cues that could hint at the original word, such as phrases that make some real-world pairings evident (e.g., 'at a [CHIMERA] angle' indicates that the original word could be 'alligator' because of the alliteration in 'alligator angle').

2. The high-FMC pseudoword's base word (e.g., 'alligator') should fit the contexts better than the medium-FMC base word (e.g., 'monkey'). For example, for the 'alligator/rattlesnake' chimera, the high-FMC base word 'alligator' should align seamlessly with all ten sentences, whereas the medium-FMC base word 'monkey' should be less congruent but still plausible. An example eligible sentence is 'But the kangaroo rat can hear the faint rustles of the [CHIMERA]'s scales moving over the sand, and escape.' It is possible for a fictional creature to resemble a monkey and have scales, even though it is much less natural than an alligator-like creature to have scales.

Ensuring the context is less congruent with the medium-FMC base word than the high-FMC base word is an intentional design. One may argue that varying the context fitness across FMC levels may introduce confounding effects of contextual congruency. However, the nature of the interference effect in lower FMC levels is exactly the discrepancy between context-based meaning inference and form-based meaning inference. For example, cognitive dissonance arises when the context hints at a land-dwelling reptile, but the word form 'morkey,' resembling 'monkey,' suggests that the creature may be related to 'monkey.' If the context fits 'monkey' equally well as

‘alligator,’ this discrepancy—and the resulting cognitive dissonance—will not occur. Nevertheless, varying the proportion of equally fitting sentences may affect the degree of cognitive dissonance and learning outcomes. In a future study, I will manipulate the level of context fitness to directly test this hypothesis, which is beyond the scope of the present study.

I edited the selected sentences to meet four additional criteria:

1. They do not contain probes or base words used for *any* chimera.
2. They do not explicitly categorize the chimera into specific categories, such as ‘vegetable’ or ‘bird.’
3. They do not mention well-known individuals or institutions.
4. They do not have grammatical errors.

By adhering to these guidelines, the selected sentences aim to provide a neutral and unbiased context across chimeras for evaluating form-meaning consistency. As an example, the reading material for the ‘alligator/rattlesnake’ chimera is presented below, using ***burket*** (bolded and italicized in the experiment) as the pseudoword. The original words in these sentences from the BNC and ukWaC corpora are presented at the end of each sentence below but did not appear in the experiment. The reading materials for all 12 chimeras can be found in the Supplemental Material: Experimental Stimuli.

He said Albert reacted like any ***burket*** with live prey, drowning it first and eating it later. [alligator]

But the kangaroo rat can hear the faint rustles of the ***burket***’s scales moving over the sand, and escape. [rattlesnake]

Large numbers of ***burket*** skins are exported to Latin America to be made into handbags, shoes and watch straps. [alligator]

The fangs of this ***burket*** are clearly visible but are not yet in the full striking position. [rattlesnake]

A widow whose arm was bitten off by a ***burket*** said yesterday she was sorry the creature was later killed. [alligator]

Burkets eat animals such as mice and the young of prairie dogs or cottontail rabbits. [rattlesnake]

Below it, the greenish water foamed over rocks and there were ***burkets*** lurking in the stony caves along the bank. [alligator]

Mulder's computer display shows a video of some evil looking hissing ***burket*** from some animal fact-type website. [rattlesnake]

The ***burket*** manages to capsize the boat but while Culp disappears beneath the water, Blackmer swims for the surface. [alligator]

There are a continually galloping rider and a ***burket*** wriggling forwards in the sand that seems to prefigure its destiny. [rattlesnake]

Probes

In this study, participants assessed the semantic relatedness between each chimera and three probe words, selected from Lazaridou et al.'s (2017) six probes. For instance, the probes for the 'alligator/rattlesnake' chimera include 'crocodile' (highest relatedness), 'iguana,' 'gorilla,' 'buzzard,' 'banner,' and 'shovel' (lowest relatedness). Adult native English speakers rated the semantic relatedness between each chimera component and each probe on a scale of 1 ('completely unrelated') to 7 ('almost the same meaning'). Each participant rated only one component, with ten ratings per component-probe pair. For instance, ten people rated the 'alligator-crocodile' pair, while another ten rated the 'rattlesnake-crocodile' pair. These ratings were averaged to establish a so-called *ground-truth chimera-probe relatedness* (CPR) as a benchmark for the learning tasks.

In line with the experimental design of three FMC conditions, I selected three probes for each chimera based on their CPR scores:

1. A probe with a high CPR score (~3-5), usually a close co-hyponym with one component of the chimera (e.g., 'crocodile' for the 'alligator/rattlesnake' chimera).
2. A probe with a medium CPR score (~2), sharing a broad category with the chimera but differing in features (e.g., 'gorilla').
3. A probe with a low CPR score (~1), unrelated to the chimera (e.g., 'shovel').

Note that Lazaridou et al. (2017) reported both word-based and image-based CPR. The image-based CPR was obtained by presenting participants with pictures of the chimera component and probes, instead of words. The two types of CPR scores generally agree with each other. Because the tasks in the present study were word-based, I used the word-based CPR for probe selection.

I excluded words with a SUBTLEXus frequency of <2 per million to avoid the effect of low word frequency on response time. Homonyms (e.g., 'crane') were also excluded. I prioritized matching word frequencies among the probes within each chimera whenever possible. Chimeras lacking a suitable set of three probes across these criteria above were excluded. The complete list of the resulting 12 chimeras I selected for the present study and their probe words can be found in the Supplemental Material: Experimental Stimuli.

Pseudowords

For each chimera, three pseudowords were generated by substituting a single letter near the center of a corresponding base word, to resemble real-word pairs like 'leopard'-'leotard':

1. **High-FMC base word**: One of the chimera components, which is a semantic neighbor of the high-relatedness probe by design (e.g., the high-FMC base word 'alligator' is a semantic neighbor of the high-relatedness probe 'crocodile')

2. **Medium-FMC base word**: A semantic neighbor of the medium-relatedness probe (e.g., the medium-FMC base word 'monkey' is a semantic neighbor of the medium-relatedness probe 'gorilla')

3. **Low-FMC base word**: a semantic neighbor of the low-relatedness probe (e.g., the low-FMC base word 'bucket' is a semantic neighbor of the low-relatedness probe 'shovel')

I operationalized semantic neighbors as words fulfilling any of the following criteria:

1. Synonyms, close hypernyms, and co-hyponyms in WordNet (Princeton University, 2010).

2. The top ten words that share semantic features with the target word, according to McRae et al. (2005) norms.

3. The top ten words commonly used in similar contexts, as calculated by the word2vec distributional semantic model using the Word Embedding Analysis tool available at http://wordvec.colorado.edu.

Additional criteria for base word selection were as follows:

1. All base words must have at least five letters. The purpose was to ensure that each base word is the sole orthographic neighbor of its corresponding pseudoword, guaranteeing that it is the most strongly activated real word when participants encounter the pseudoword. When base words have four or fewer letters, generating a pseudoword that fulfills this criterion is nearly impossible. For the high-FMC base word, if neither chimera component meets this criterion, a semantic neighbor is used (e.g., 'pumpkin' is chosen for the 'corn/yam' chimera).

2. Semantic neighbors with a SUBTLEXus frequency of < 2 per million were excluded. The final base words for each chimera were selected to have similar frequencies and lengths.

3. Words already selected as probes for semantic relatedness tasks were excluded. Each base word was used only once across all chimeras.

Each pseudoword has one and only one real-word orthographic neighbor, which is the base word. Table 2 illustrates the pseudoword generation for the 'alligator/rattlesnake' chimera.

Table 2

Pseudoword Generation Method

Probe	Base word	Pseudoword
crocodile	alligator	alli**b**ator
gorilla	monkey	mo**r**key
shovel	bucket	bu**r**ket

For the control condition, a pseudoword with no real-word orthographic neighbors was generated to match the average lengths of the other three pseudowords.

All pseudowords were generated using WordGen (Duyck et al., 2004), adhering to the default criteria that produce pseudowords with high legality: 1. Minimum legal bigram frequency > 30; 2. Minimum position-specific onset/suffix bigram frequency > 15. In this context, bigrams refer to adjacent letter pairs in a word. The onset bigram is the first letter pair, while the suffix bigram is the last pair. For example, the pseudoword 'burket' comprises the following five bigrams, with their respective non-position-specific frequencies (per million words) in parentheses: bu (onset, 583), ur (1573), rk (365), ke (1040), et (suffix, 1836).

Participants

I recruited adult, monolingual native English speakers through Prolific's online crowdsourcing platform (www.prolific.co). The experiment was conducted online using the Gorilla software (www.gorilla.sc). The inclusion criteria were current U.S. resident, 18-40 years old, monolingual native English speaker, no diagnosed language or reading impairments, and normal or corrected-to-normal vision. This research received ethical approval from UC Berkeley's Committee for the Protection of Human Subjects (Protocol #2023-07-16544).

Procedure

A practice trial preceded the main tasks. Participants were asked to read ten sentences containing a pseudoword 'gaddil,' a chimera of 'trousers' and 'shirt.' They were given the following instructions prior to the reading:

'You will read a set of ten sentences featuring a new word (***italicized and bolded***). This word may look like a real English word, but it is an entirely new word. It refers to an imaginary entity that could belong in a science fiction universe.

Please try to **learn and remember the new word and its potential meaning**. You will be tested on your knowledge of this new word after reading the sentences.

You will need to **read all sentences carefully** to learn as much as possible about the new word.

Please do not write anything down.'

The reading was followed by the semantic relatedness rating task, with the following instructions:

'You will be presented with the new word you just saw in the reading, alongside a familiar word.

Please rate the relatedness between these two words, focusing solely on **meaning** and not on spelling or sound.

The relation could be stricter or looser. Please rate the relatedness on a 5-point scale (1 = unrelated at all in meaning, 5 = almost the same meaning).

The **new word** will always appear on **the left side of the screen**.

Press the number on your keyboard **as fast as you can while remaining accurate**.'

Before the rating, a separate screen was shown: 'Please remember to **respond as quickly as possible while remaining accurate**' in large bolded font. They were then asked to rate the relatedness between 'gaddil' and three probes, 'pajama,' 'curtain,' 'apricot,' one at a time in random order. After the rating, another separate screen was shown: 'In the main experiment that comes next, please also be sure to make your ratings **as quickly as possible while remaining accurate.**'

The main experiment started after the practice trial. For each participant, the order of 12 experimental reading materials was randomized and then divided into three blocks. Each block consisted of four reading materials, one for each chimera.

Participants were randomly assigned to either the delayed-only (23 participants) or immediate+delayed (27 participants) testing group. For participants in the delayed-only group, after reading the materials in each block, they engaged in a 1-minute filler task (detailed below), and then performed the relatedness judgment tasks for the four pseudowords they had just encountered. The 3×4=12 pseudoword-probe pairings were presented individually and in a fully randomized order for each participant. The participants in the immediate+delayed group performed rating tasks immediately after reading *each* passage, in addition to the delayed rating tasks.

Each filler task consisted of three trials where participants were asked to count the number of zeros in a 5×5 table with varying numbers of ones and zeros on each row. This filler-task design had two purposes. Firstly, it could mitigate the recency effect for the most recently encountered words. Secondly, it served as an attention check. Participants had three chances to enter the correct answer for each trial; failing three times led to a warning message and they had to wait for 30 seconds to try again. The exclusion criteria were receiving the warning message more than once. All participants passed the attention check.

The immediate+delayed group's performance in the immediate rating task would help us understand if the effect of form-meaning consistency already occurs at the stage of inferring the meaning during reading and immediate retrieval after reading (e.g., Does one assume that 'burket' may be related to 'bucket' when trying to infer what 'burket' means during and immediately after reading?).

On the other hand, the two groups' performance in the delayed test would shed light on to what extent the effect of form-meaning consistency also occurs at the stage of retrieving the meaning later (e.g., When trying to recall what 'burket' means several minutes after reading, does the activation of 'bucket' interfere with retrieving the correct meaning?).

We asked the immediate+delayed group to perform both the immediate and delayed tasks so that we could compare the delayed task performance between the delayed-only group and the immediate+delayed group. If the hypothesized interference effects of lower FMC in the delayed tasks were attenuated in the immediate+delayed group in comparison to the delayed-only group,

it would indicate that performing semantic tasks immediately after learning could reduce the interference in later retrieval, which has pedagogical implications.

At the experiment's conclusion, participants were asked whether they took notes during the session while being assured that admitting to note-taking would not affect their compensation. Their data would nonetheless be excluded if they confirmed that they took notes. No participants reported taking notes. Additionally, individual data for a specific chimera would be omitted if the time spent reading the corresponding material fell 3 standard deviations *below* the mean. No observations fell into this range. The reading times for six reading materials (across five participants) in the delayed-only group and six reading materials (across three participants) in the immediate+delayed group fell 3 standard deviations *above* the mean, ranging from 142 seconds to 305 seconds, which were reasonable for reading ten sentences while trying to infer the meaning of a novel word and thus did not indicate that the participants were distracted. Spending too little time, instead of too much time, was more of a problem in this scenario because the participants may not have put in effort to learn as much about the word as possible.

Instead of excluding items whose response times deviated by more than 3 standard deviations from the mean, I followed the practice of Forster and Hector (2002) to recode the outliers. I trimmed the item response times for each participant using a threshold of 2 standard deviations above and below the mean for that specific participant. Response times outside these thresholds were recoded to the threshold value, instead of being excluded. Forster and Hector (2002) argued that this method of preprocessing is more conservative for experiments that aim to explore inhibitory effects; additionally, disregarding longer RTs entirely could lead to an underestimation of the inhibitory effects.

The programmed experiment is publicly available on app.gorilla.sc/openmaterials/774924.

Task-Specific Hypotheses

I propose the following conceptual model (Figure 4) to explain how Form-Meaning Consistency (FMC) impacts semantic relatedness ratings. In this model:

The context-based response represents the response based on meaning inference from contextual clues only, corresponding to the intended relatedness between the chimera and the probe. For instance, based on the context, 'burket' refers to a land-dwelling reptile and has 'low' semantic relatedness with 'shovel,' 'medium' relatedness with 'gorilla,' and 'high' relatedness with 'crocodile.'

The form-based response represents the response based on meaning inference from word form only, corresponding to the relatedness between the activated real word and the probe. For instance, based on the word form, 'burket' resembles 'bucket,' which may indicate that 'burket' could be related to 'bucket,' which has 'high' semantic relatedness with 'shovel,' and 'low' relatedness with both 'gorilla' and 'crocodile.'

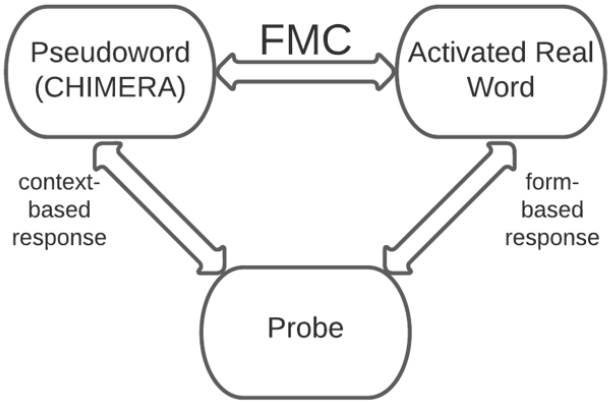
I hypothesize that in a semantic relatedness task, when one sees a pseudoword resembling a real word, the real word will be activated, and a rapid form-based response will be made, in addition to the context-based response. Discrepancies between the context-based and form-based responses can lead to upward or downward biases in ratings and longer response times. The

discrepancies and lack thereof are represented as ‘<,’ ‘>,’ ‘=’ in the center of Figures 4a-4d. For instance, in the ‘burket-shovel’ rating, the context-based response (low, on the left side) is lower than (represented as ‘<’) the form-based response (high, on the right side).

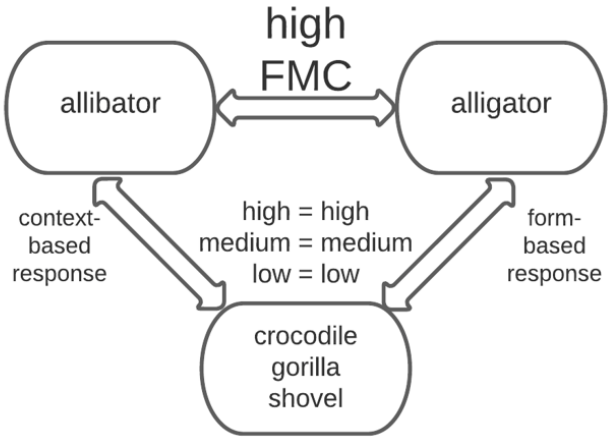
Figure 4

Form-Meaning Consistency, Chimera-Probe Relatedness, and Real Word-Probe Relatedness

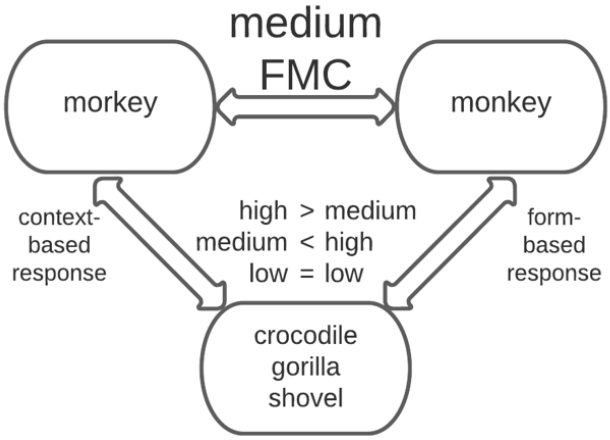
4a



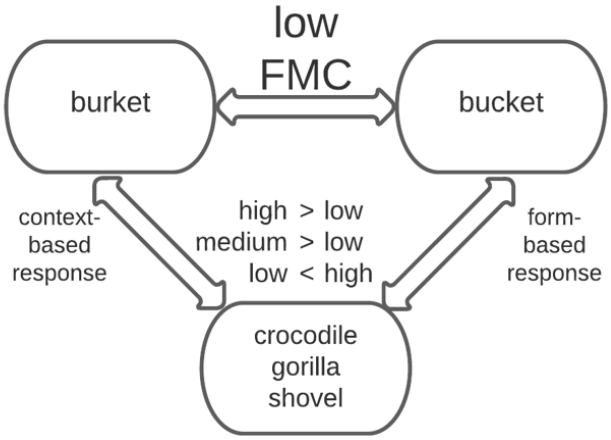
4b



4c



4d



In cases of high FMC, for all probes, there is no discrepancy between the context-based and form-based responses (illustrated by the three '=' symbols in Figure 4b), so the final rating should not be biased. The response time for high FMC will be shorter than all other conditions for two reasons: 1. The activated real word is semantically related to the pseudoword; it is, in fact, one component of the chimera, thus enabling quicker retrieval of the correct semantic representation. 2. There is no discrepancy between the form-based and context-based response, so decision-making will not be delayed.

In medium or low FMC cases, there could be discrepancies between the context-based and form-based responses:

1. When the form-based response for semantic relatedness is lower than the intended context-based response, the final rating may have a downward bias. For instance, in

Figure 4d, the form-based response of the relatedness between 'bucket' and 'crocodile' is low, which is lower than the context-based response for rating 'burket' against 'crocodile' (high), thus potentially leading to a downward bias towards 'low.'

2. When the form-based response is higher than the context-based response, the final rating may have an upward bias. For instance, in Figure 4d, the form-based response of the relatedness between 'bucket' and 'shovel' is high, which is higher than the context-based response for rating 'burket' against 'shovel' (low), thus potentially leading to an upward bias towards 'high.'

In other words, the final ratings may be biased towards the form-based response whenever a discrepancy occurs; I expect this bias to be subtle but systematic. Either case of response discrepancy may lead to a longer response time. Additionally, the lower the FMC is, the longer the response time may be, above and beyond the effect of response discrepancy. The reason is that more cognitive resources are needed to suppress the irrelevant semantic information and retrieve the correct meaning when the activated real word is less semantically related to the pseudoword.

In a nutshell, lower degrees of FMC (e.g., 'burket') may lead to different types of discrepancy between the context-based response (e.g., high, the true relatedness between 'burket' and 'crocodile') and the form-based response (e.g., low, the relatedness between 'bucket' and 'crocodile'), which, in turn, may lead to overestimation or underestimation. The higher the discrepancy is, the larger the bias may be. On the other hand, the delay in response time under lower FMC can be caused by two reasons:

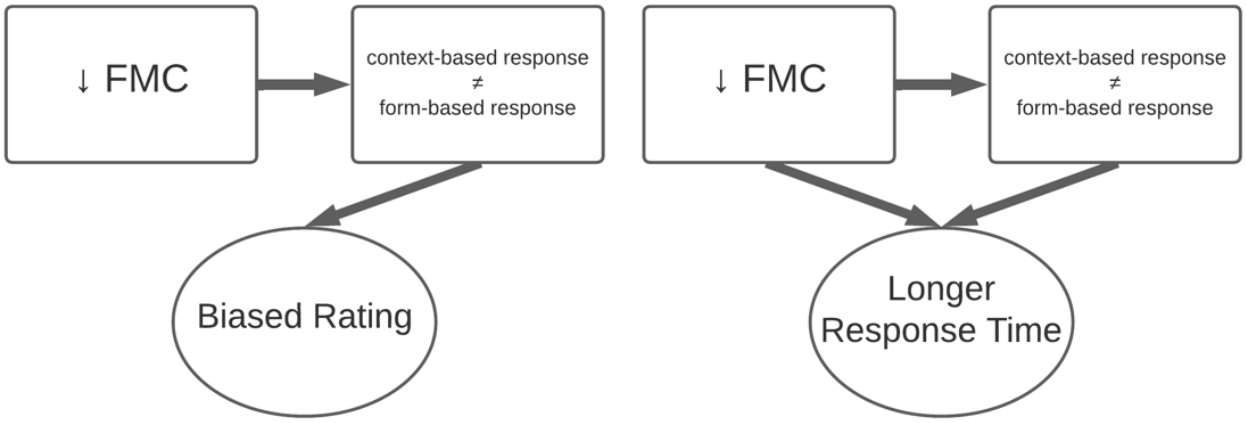
1. One needs to suppress the unrelated meaning of the activated real word (e.g., 'bucket'), delaying the retrieval of the correct semantic representation (e.g., alligator-like animal), regardless of the existence of response discrepancy; in contrast, high FMC (e.g., 'allibator') leads to the activation of the semantically related real-word neighbor (e.g., 'alligator'), which makes the retrieval of the correct meaning easier.

2. One also needs to suppress the incorrect form-based response before settling on the correct context-based response because of the task-specific discrepancy between these two types of responses. This additional cognitive dissonance further delays decision-making, which is similar to response inhibition (MacLeod, 1991) in incongruent Stroop tasks (e.g., name the ink color for the word 'blue' printed in red ink) where participants have to suppress their incorrect response based on the word's meaning before making the correct response based on the ink color.

These pathways of how FMC and the discrepancy between the context-based and form-based responses can affect task performance are illustrated in Figure 5.

Figure 5

The Effect of Form-Meaning Consistency and Discrepancy Between Context-Based and Form-Based Responses on Rating Scores and Response Time



The influence of FMC and response discrepancy on participants' performance would be attenuated in the immediate ratings because of the low demand on memory retrieval. It would be more salient in the delayed rating. The reason is that in the delayed rating, the participants need to first recognize the pseudoword's form and then retrieve the meaning associated with that word form; this process may be more vulnerable to the activation of the known real word, especially at this early stage of learning where the connection between the new word form and meaning is still weak. This may be especially the case for participants in the delayed-only group but attenuated for those in the immediate+delayed group. When participants in the immediate+delayed group perform the delayed rating, their immediate rating may still be fresh in the memory and thus may influence their delayed rating; additionally, the immediate rating task may have consolidated their memory of the pseudoword's form and meaning and thus they may be less swayed by the real word activated. That being said, we predict that the effect of FMC would still be observed in the delayed response for the immediate+delayed group, even if it may not be as salient as the delayed-only group.

Accordingly, I would make the following comparisons and test the specific hypotheses on rating score and response time, as shown in Table 3.

Table 3

Comparisons and Hypotheses for Relatedness Tasks

Types of Comparisons		Ratings	Response Time
Same FMC condition against different probes			
High FMC	allibator-crocodile vs. allibator-gorilla vs. allibator-shovel	High Probe > Medium Probe > Low Probe	Same across probes
Medium FMC	morkey-crocodile vs. morkey-gorilla vs. morkey-shovel	High / Medium Probe > Low Probe	High / Medium Probe > Low Probe
Low FMC	burket-crocodile vs. burket-gorilla vs. burket-shovel	High probe > Medium / Low Probe	Same across probes
Same probe against different FMC conditions			
Low-relatedness probe	allibator-shovel vs. morkey-shovel vs. burket-shovel vs. darane-shovel	Low FMC > High / Medium FMC / Control	Low FMC > Medium FMC > Control > High FMC
Medium-relatedness probe	allibator-gorilla vs. morkey-gorilla vs. burket-gorilla vs. darane-gorilla	Medium FMC > High FMC / Control > Low FMC	Low FMC > Medium FMC > Control > High FMC
High-relatedness probe	allibator-crocodile vs. morkey-crocodile vs. burket-crocodile vs. darane-crocodile	High FMC / Control > Medium FMC > Low FMC	Low FMC > Medium FMC > Control > High FMC

Statistical Analyses

To analyze the rating scores obtained from the semantic relatedness tasks, I employ Linear Mixed-Effects Models (LMMs). The fixed effects in the model include Form-Meaning Consistency (FMC) level with four levels (High, Medium, Low, and Control), Probe Type with three levels (High, Medium, and Low CPR), and their interactions. I also included 'Item Order', representing the order of the three items within each chimera, as a predictor to control for within-task learning effects. Random effects of participants and chimeras (nested within participants) accounted for individual variability across participants and chimeras. The LMM formula for rating scores is (using syntax for the R-package lme4):

$$\text{Rating Score} \sim \text{FMC Level} \times \text{Probe Type} + \text{Item Order} + (1|\text{Participant: Chimera})$$

For analyzing response times in semantic relatedness tasks, I will also use LMMs. The model will include the same fixed and random effects as the model for rating scores.

All models were fitted using the 'mixed' command in Stata. Each of these models has been selected to best capture the data's structure and allow for the most accurate interpretation of the effects of interest (as detailed in the **Task-Specific Hypotheses** section). Model assumptions on residuals were checked and no violation of the normality assumption was identified (see Supplemental Material: Model Diagnostics).

CHAPTER III: RESULTS

Table 4 presents the model estimates of six models: 1. Ratings for the delayed-only group; 2. Ratings for the immediate + delayed (IM+DL) group's immediate tasks; 3. Ratings for the IM+DL group's delayed tasks; 4. Response time (RT) for the delayed-only group; 5. RT for the IM+DL group's immediate tasks; 6. RT for the IM+DL group's delayed tasks.

Table 4

Model Estimates

	(1) Rating - Delayed Only	(2) Rating - IM + DL Immediate	(3) Rating - IM + DL Delayed	(4) RT - Delayed Only	(5) RT - IM + DL Immediate	(6) RT - IM + DL Delayed
Medium FMC vs. Low FMC	-0.566**	-0.047	-0.251	-655.692***	-107.914	1.145
	(0.206)	(0.153)	(0.163)	(178.823)	(76.320)	(105.914)
High FMC vs. Low FMC	-0.461*	-0.124	-0.300	-614.243***	-93.475	-340.525**
	(0.206)	(0.153)	(0.163)	(179.045)	(76.246)	(105.849)
Control vs. Low FMC	-0.284	0.005	-0.189	-192.735	-33.425	-74.998
	(0.206)	(0.153)	(0.163)	(179.012)	(76.287)	(105.916)
Medium Probe vs. Low Probe	0.330	0.923***	0.795***	99.566	282.367***	100.902
	(0.207)	(0.153)	(0.164)	(167.041)	(73.591)	(104.637)
High Probe vs. Low Probe	1.560***	2.664***	2.594***	66.806	367.809***	133.985
	(0.206)	(0.153)	(0.163)	(166.577)	(73.569)	(104.257)
Medium FMC×Medium Probe	1.032***	0.398	0.670**	447.898	116.906	-37.498
	(0.291)	(0.217)	(0.232)	(235.479)	(104.090)	(148.678)
Medium FMC×High Probe	0.710*	-0.059	0.232	345.712	198.107	-156.518
	(0.291)	(0.217)	(0.230)	(235.591)	(104.092)	(147.379)
High FMC×Medium Probe	0.572*	-0.002	0.526*	312.807	117.676	183.321
	(0.292)	(0.216)	(0.231)	(236.048)	(103.928)	(147.970)
High FMC×High Probe	0.854**	-0.032	0.202	359.956	116.728	228.426
	(0.292)	(0.217)	(0.231)	(235.710)	(104.001)	(147.696)
Control×Medium Probe	0.495	-0.083	-0.077	207.399	13.614	-61.555
	(0.292)	(0.217)	(0.232)	(235.966)	(104.131)	(148.224)
Control×High Probe	0.268	0.304	0.176	23.987	-16.355	53.569
	(0.292)	(0.217)	(0.230)	(235.694)	(103.947)	(147.390)
Within Chimera Order 2 vs. 1	0.016	0.064	-0.049	-522.617***	-234.869***	-325.245***
	(0.103)	(0.077)	(0.083)	(83.596)	(37.197)	(52.777)
Within Chimera Order 3 vs. 1	-0.087	0.108	-0.026	-636.387***	-273.439***	-287.182***
	(0.103)	(0.077)	(0.082)	(83.663)	(37.065)	(52.594)

Intercept	2.060***	1.162***	1.496***	2986.236***	1434.039***	1937.990***
	(0.186)	(0.129)	(0.132)	(245.640)	(81.628)	(118.737)
Ln(Standard Deviation of Person Random Effects)	-0.745***	-1.260***	-1.569***	6.890***	5.696***	6.108***
	(0.174)	(0.181)	(0.230)	(0.153)	(0.148)	(0.145)
Ln(Standard Deviation of Chimera Random Effects)	-14.437	-17.397	-17.966	5.948***	4.860***	4.749***
	(1395.745)	(813.569)	(815.561)	(0.155)	(0.252)	(0.579)
Ln(Standard Deviation of Item Residuals)	0.190***	-0.026	0.036	6.885***	6.148***	6.497***
	(0.025)	(0.023)	(0.023)	(0.030)	(0.028)	(0.028)
N	828	972	972	828	972	972

Note. Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

I will examine whether the model estimates support or challenge the task-specific hypotheses presented in Table 3, one at a time, including:

1. High FMC against different probes: a. ratings; b. response time
2. Medium FMC against different probes: a. ratings; b. response time
3. Low FMC against different probes: a. ratings; b. response time
4. Low probe against different FMC conditions: a. ratings; b. response time
5. Medium probe against different FMC conditions: a. ratings; b. response time
6. High probe against different FMC conditions: a. ratings; b. response time

Under each type of comparison, I will first discuss the specific hypotheses I had made before collecting the data and the reason behind these hypotheses. I will then present the model estimates and discuss whether they supported or challenged the hypotheses and why. The results and discussions for the delayed-only group will be presented first, followed by the immediate task and then the delayed task of the immediate+delayed group. This chapter will end with an overall summary of the results and discussions.

Even though this chapter's title is 'Results,' this chapter will be a combination of results and detailed discussion of these results, due to the sheer number of task-specific hypotheses being examined. The next chapter, 'Discussion,' will be a reorganization of the discussion in this chapter to examine broader hypotheses about the mechanisms of how FMC influences task performances.

High FMC Against Different Probes

Table 5 juxtaposes the hypotheses and the model estimates for comparing low, medium, and high probes within the high FMC condition. The hypotheses and model estimates for ratings and response time are on separate rows. The model estimates for the delayed-only group, the immediate tasks for the IM+DL group, and the delayed tasks for the IM+DL group are in separate columns. For the model estimates, the low probe was set as the reference group so that the point estimates of high probes and medium probes represent the difference in estimated mean ratings or

response times between that type of probes and low probes, within the high FMC condition. The p values and standardized effect sizes (es; the result of dividing the estimated coefficient by the overall standard deviation of the outcome variable) for each comparison are also provided next to the point estimates.

Figure 6a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 6b shows the model estimated mean ratings for low, medium, and high probes within the high FMC condition, across the three testing groups, setting within-chimera item order to 1. Figure 6c is the same type of graph for the model estimated mean response times.

Table 5

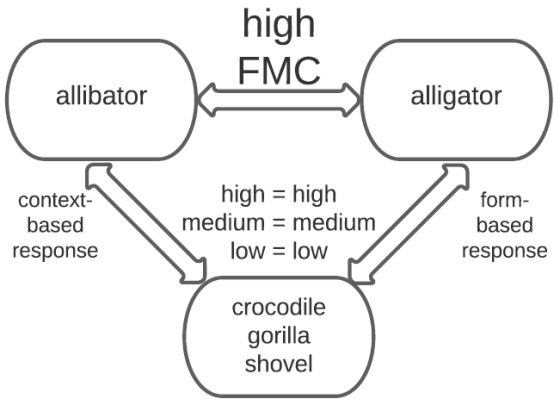
High FMC Against Different Probes: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
High FMC	allibator-crocodile allibator-gorilla allibator-shovel	**Ratings:** High Probe > Medium Probe > Low Probe	**High Probe** **(2.414, p < 0.001, es = 1.555)** **Medium Probe** **(0.903, p < 0.001, es = 0.581)** *Low Probe* *(Reference)*	**High Probe** **(2.632, p < 0.001, es = 1.733)** **Medium Probe** **(0.921, p < 0.001, es = 0.606)** *Low Probe* *(Reference)*	**High Probe** **(2.796, p < 0.001, es = 1.797)** **Medium Probe** **(1.321, p < 0.001, es = 0.849)** *Low Probe* *(Reference)*
		Response Time: Same across probes	**High Probe** **(426.762, p = 0.010, es = 0.287)** **Medium Probe** **(412.373, p = 0.013, es = 0.278)** *Low Probe* *(Reference)*	**High Probe** **(484.537, p < 0.001, es = 0.788)** **Medium Probe** **(400.043, p < 0.001, es = 0.650)** *Low Probe* *(Reference)*	**High Probe** **(362.411, p = 0.001, es = 0.437)** **Medium Probe** **(284.222, p = 0.006, es = 0.342)** *Low Probe* *(Reference)*

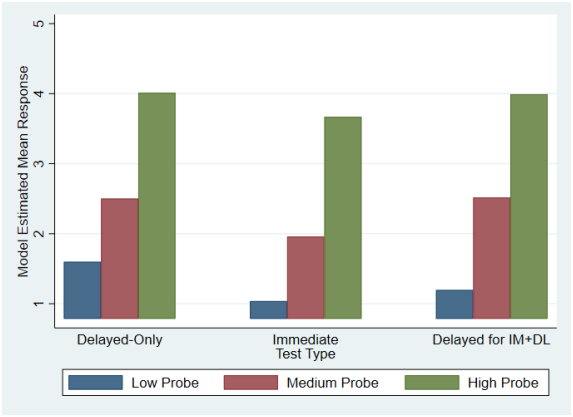
Figure 6

High FMC Against Different Probes: Response Discrepancies and Model Estimates

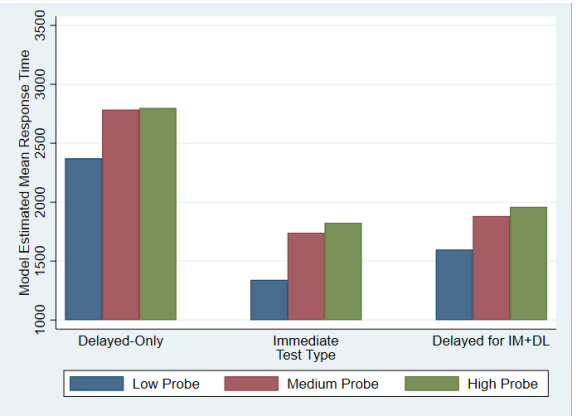
6a



6b



6c



Ratings

Hypothesis. For the high FMC condition (e.g., 'allibator' for the alligator/rattlesnake chimera), I predicted that no rating bias would be observed for any of the three types of probes because there would be no discrepancy between the context-based and form-based responses (see Figure 6a). Thus, the high probe (e.g., 'crocodile') rating should be higher than the medium probe (e.g., 'gorilla'), which would be higher than the low probe (e.g., 'shovel').

Delayed-Only Group. For the delayed-only group, model estimates supported this hypothesis, indicating significant differences with large standardized effect sizes across all three pairs of comparisons (see Figure 6b): 1. High probe > Low probe (2.414 points, $p < 0.001$, es = 1.555); 2. Medium probe > Low probe (0.903 points, $p < 0.001$, es = 0.581), 3. High probe > Medium probe (1.511 points, $p < 0.001$, es = 0.974).

Immediate Tasks for IM+DL Group. For the immediate response of the immediate+delayed group, the differences across the three pairs of comparisons were also all significant with large standardized effect size (see Figure 6b): 1. High Probe > Low Probe (2.632 points, $p < 0.001$, es = 1.733); 2. Medium probe > Low probe (0.921 points, $p < 0.001$, es = 0.606); 3. High probe > Medium probe (1.711 points, $p < 0.001$, es = 1.127).

Delayed Tasks for IM+DL Group. The same trend was observed for the delayed tasks of the immediate+delayed group as well (see Figure 6b): 1. High Probe > Low Probe (2.796 points, $p < 0.001$, es = 1.797); 2. Medium probe > Low probe (1.321 points, $p < 0.001$, es = 0.849); 3. High probe > Medium probe (1.475 points, $p < 0.001$, es = 0.948).

Response Time

Hypothesis. For the high FMC condition, I predicted that the response time would be the same across all types of probes, because there would be no discrepancies between the context-based and form-based responses across the three probes.

Delayed-Only Group. Surprisingly, for the delayed-only group, model estimates indicated that the mean response time for high probes was estimated to be 426.762 ms higher than low probes ($p = 0.010$, es = 0.287); the mean response time for medium probes was estimated to be 412.373 ms higher than low probe ($p = 0.013$, es = 0.278). The mean response time was not significantly different between high probes and medium probes (14.389 ms, $p = 0.931$, es = 0.009). This result indicates that low-relatedness probes were generally faster to rate than higher-relatedness probes (see Figure 6c). One would quickly press '1' whenever they felt that the two words were completely unrelated (e.g., 'alligator' vs. 'crocodile'), whereas higher relatedness probes would take slightly more time to decide exactly which number was the most appropriate. One may be indecisive between 2-3 for a medium-relatedness probe and between 4-5 for a high-relatedness probe. For example, one may wonder, 'Is the relatedness between an alligator-like chimera and gorilla a 2 or a 3?'

Immediate Tasks for IM+DL Group. A similar trend has been observed in the immediate task for the immediate+delayed group (see Figure 6c): 1. High Probe > Low Probe (484.537 ms, $p < 0.001$, es = 0.788); 2. Medium Probe > Low Probe (400.043, $p < 0.001$, es = 0.650); 3. High Probe $\approx$ Medium Probe (84.494 ms, $p = 0.252$, es = 0.138).

Delayed Tasks for IM+DL Group. A similar trend has also been observed in the delayed task for the immediate + delayed group (see Figure 6c): 1. High Probe > Low Probe (362.411 ms, $p = 0.001$, es = 0.437); 2. Medium Probe > Low Probe (284.222, $p = 0.006$, es = 0.342); 3. High Probe $\approx$ Medium Probe (78.188 ms, $p = 0.456$, es = 0.095). These results further corroborated the explanation above regarding the lower decision time for low-relatedness probes.

Medium FMC Against Different Probes

Table 6 juxtaposes the hypotheses and the model estimates for comparing low, medium, and high probes within the medium FMC condition. Figure 7a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 7b shows the model estimated mean ratings for low, medium, and high probes within the medium FMC condition, across the three testing groups, setting within-chimera item order to 1. Figure 7c is the same type of graph for the model estimated mean response times.

Table 6

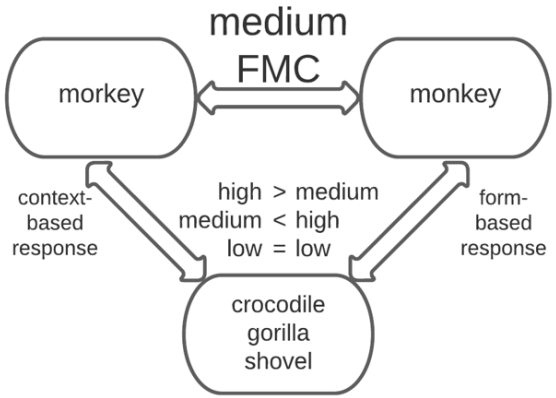
Medium FMC Against Different Probes: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
Medium FMC	morkey-crocodile morkey-gorilla morkey-shovel	**Ratings:** High / Medium Probe > Low Probe	**High Probe** **(2.270, p < 0.001, es = 1.462)** **Medium Probe** **(1.363, p < 0.001, es = 0.878)** *Low Probe* *(Reference)*	**High Probe** **(2.606, p < 0.001, es = 1.715)** **Medium Probe** **(1.321, p < 0.001, es = 0.870)** *Low Probe* *(Reference)*	**High Probe** **(2.826, p < 0.001, es = 1.816)** **Medium Probe** **(1.465, p < 0.001, es = 0.941)** *Low Probe* *(Reference)*
		Response Time: High / Medium Probe > Low Probe	**Medium Probe** **(547.464, p = 0.001, es = 0.369)** **High Probe** **(412.518, p = 0.013, es = 0.278)** *Low Probe* *(Reference)*	**High Probe** **(565.916, p < 0.001, es = 0.920)** **Medium Probe** **(399.274, p < 0.001, es = 0.649)** *Low Probe* *(Reference)*	Medium Probe (63.404, p = 0.545, es = 0.076) *Low Probe* *(Reference)* High Probe (-22.533, p = 0.829, es = -0.027)

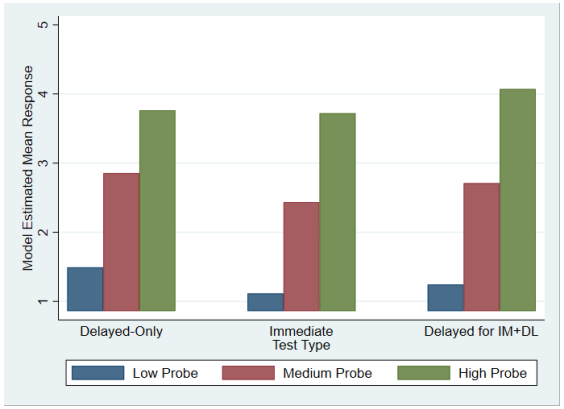
Figure 7

Medium FMC Against Different Probes: Response Discrepancies and Model Estimates

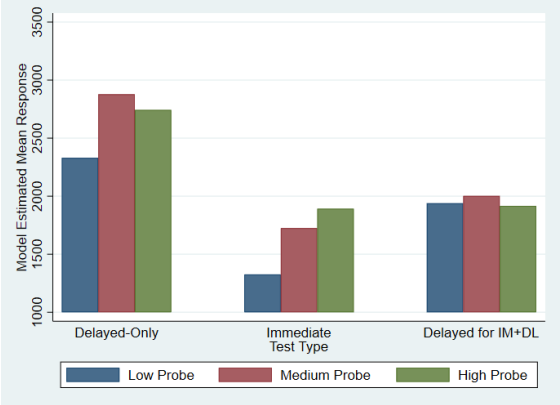
7a



7b



7c



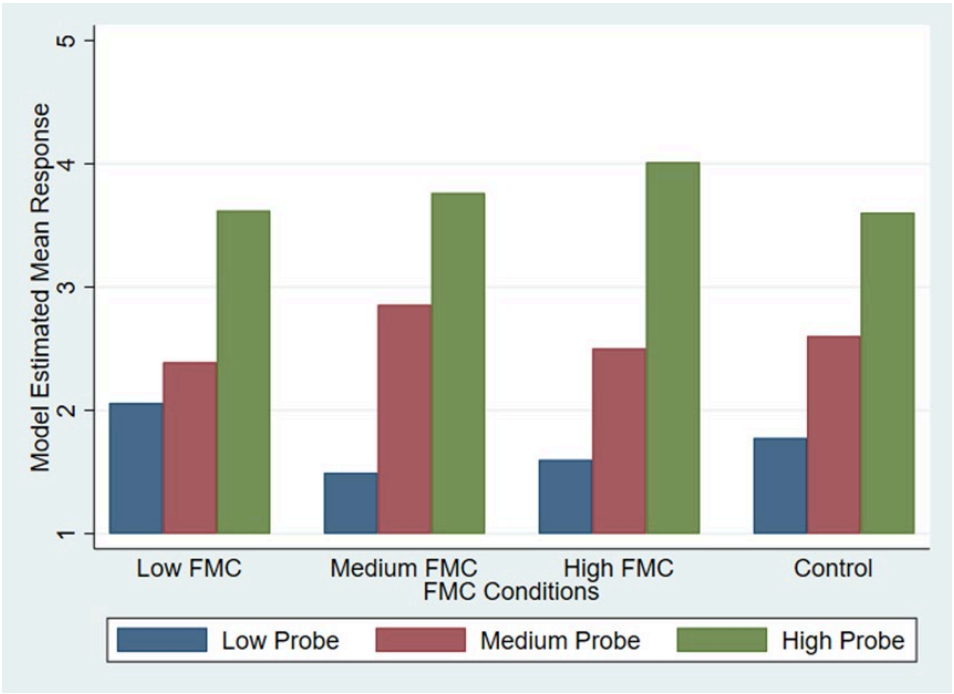
Ratings

Hypothesis. For medium FMC (e.g., 'morkey'), I predicted that the medium probe (e.g., 'gorilla') would be overestimated because the context-based response (medium) is lower than the form-based response (high; e.g., 'monkey' vs 'gorilla'), whereas the high probe (e.g., 'crocodile') would be underestimated because the context-based response (high) would be higher than the form-based response (medium; e.g., 'monkey' vs. 'crocodile'), as illustrated in Figure 7a. The underestimation of the high probe and the overestimation of the medium probe could decrease the gap between the ratings of these two probes or even render the two ratings comparable to each other. On the other hand, the ratings of low probes ('shovel') would remain unbiased because of no discrepancy between the context-based and form-based responses.

Delayed-Only Group. Model estimates for the delayed-only group confirmed these hypotheses. Even though the mean rating of high probes was still estimated to be significantly higher than medium probes, the gap between high probes and medium probes did decrease in comparison to all other FMC conditions (including control), as predicted. The high-medium probe difference was estimated to be 0.908 ($p < 0.001$, es = 0.584, see Figure 8 below) for the medium FMC condition, in comparison to 1.511 for high FMC condition ($p < 0.001$, es = 0.974, see Figure 8), 1.230 for the low FMC condition ($p < 0.001$, es = 0.792, see Figure 8), and 1.003 for the control condition ($p < 0.001$, es = 0.646, see Figure 8). Notably, the difference in mean ratings between high and medium probes was estimated to be 0.604 points higher for the high FMC condition than the medium FMC condition, which was statistically significant ($p = 0.038$, es = 0.390). This result provides evidence that the discrepancy between the context-based and form-based inferences caused by the medium FMC did lead to overestimations of medium probes and underestimations of high probes.

Figure 8

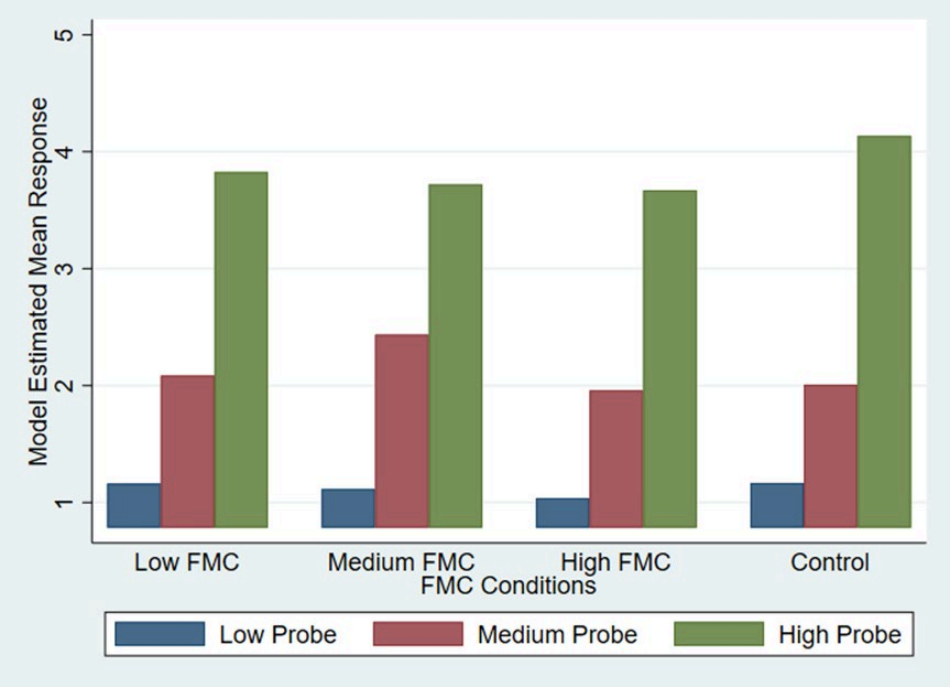
Model Estimates of Delayed-Only Group: Four Conditions × Three Probes



Immediate Tasks for IM+DL Group. The results for the immediate response were similar to the delayed-only group. Even though the mean rating of high probes was still estimated to be significantly higher than medium probes, the difference between the high probe and the medium probe was lower in the medium FMC condition than the other conditions, as predicted. The high-medium probe difference was estimated to be 1.285 ($p < 0.001$, es = 0.845, see Figure 9 below) for the medium FMC condition, in comparison to 1.711 for high FMC condition ($p < 0.001$, es = 1.127, see Figure 9), 1.741 for the low FMC condition ($p < 0.001$, es = 1.146, see Figure 9), and 2.128 for the control condition ($p < 0.001$, es = 1.401, see Figure 9). Notably, the difference in mean ratings between high and medium probes was estimated to be 0.828 points higher for the control condition than the medium FMC condition, which was statistically significant ($p = 0.016$, es = 0.556).

Figure 9

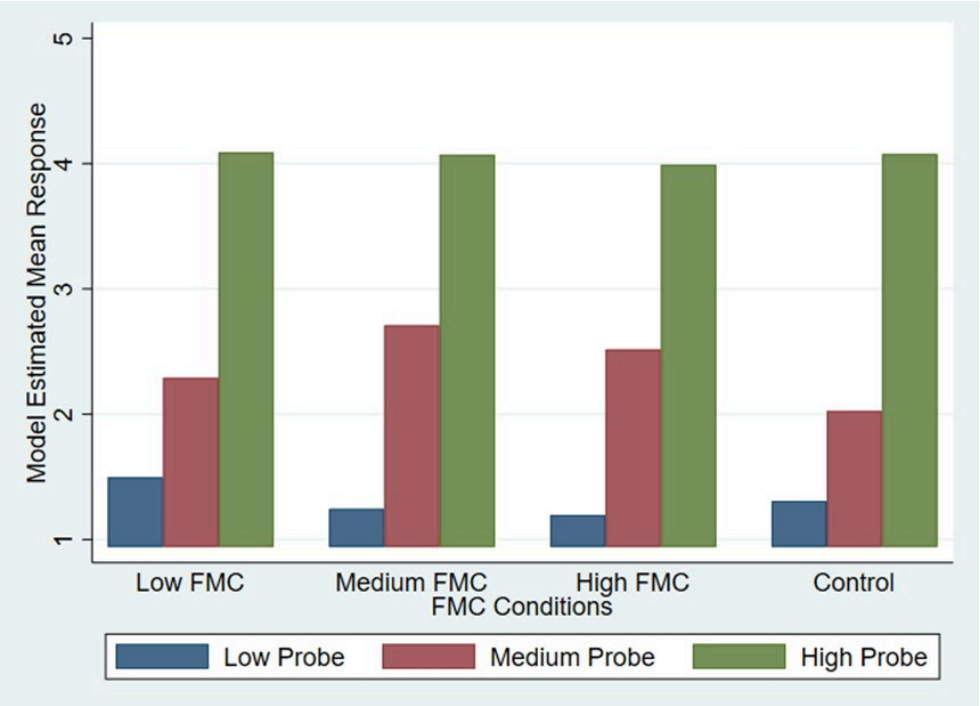
Model Estimates of Immediate+Delayed Group's Immediate Tasks: Four Conditions × Three Probes



Delayed Tasks for IM+DL Group. Similar findings have been observed for the delayed tasks of the immediate+delayed group. Even though the mean rating of high probes was still estimated to be significantly higher than medium probes, the difference between the high probe and the medium probe was lower in the medium FMC condition than other conditions, as predicted. The high-medium probe difference was estimated to be 1.361 ($p < 0.001$, es = 0.785, see Figure 10 below) for the medium FMC condition, in comparison to 1.475 for high FMC condition ($p < 0.001$, es = 0.948, see Figure 10), 1.799 for the low FMC condition ($p < 0.001$, es = 1.156, see Figure 10), and 2.051 for the control condition ($p < 0.001$, es = 1.318, see Figure 10). Notably, the difference between high and medium probes was 0.689 points higher for the control condition than the medium FMC condition, which was statistically significant ($p = 0.003$, es = 1.266).

Figure 10

Model Estimates of Immediate+Delayed Group's Delayed Tasks: Four Conditions × Three Probes



Response Time

Hypothesis. I predicted that in medium FMC condition ('morkey'), high probe (e.g., 'crocodile') and medium probe (e.g., 'shovel') would have longer response time because of the discrepancy between the context-based and form-based responses (see Figure 7a). The high probe and medium probe should be comparable to each other because the degree of the response discrepancy is the same (i.e., the discrepancy between high and medium is the same as the discrepancy between medium and high). On the other hand, the low probe had no response discrepancy, so its response time would not be affected.

Delayed-Only Group. For the delayed-only group, model estimates confirm these hypotheses, showing that the medium probes and high probes were estimated to have a mean response time of 547.464 ms ($p = 0.001$, es = 0.369) and 412.518 ms ($p = 0.013$, es = 0.278) higher than low probe, respectively (see Figure 7c). The estimated mean response times of the medium probes and high probes were comparable to each other (134.946 ms, $p = 0.419$, es = 0.091).

Immediate Tasks for IM+DL Group. Similar to the delayed-only group, for the immediate ratings of the immediate+delayed group, the medium probes and high probes were estimated to have a mean response time of 565.916 ms ($p < 0.001$, es = 0.920) and 399.274 ms ($p < 0.001$, es = 0.649) higher than low probes, respectively (see Figure 7c).

However, different from the delayed-only group, for the immediate ratings of the immediate+delayed group, the estimated mean response times of the high probes were 166.642 ms higher than the medium probes ($p = 0.024$, es = 0.271). A possible explanation is that some participants assumed that 'morkey' was related to 'monkey' when trying to infer its meaning even though the 'monkey' meaning did not fit the context perfectly. Thus, in the immediate rating, the context-based response became high instead of medium, which is the same as the form-based response (i.e., high). Hence, the medium probes, which had no response discrepancy, had lower response time than high probes, which did have response discrepancy (high > medium, see Figure 7a).

Delayed Tasks for IM+DL Group. For the delayed ratings of the immediate+delayed group, no differences were found in mean response time across all three probes (see Figure 7c). This could be because the memory of the immediate ratings was still fresh in mind so participants could make equally quick judgments across the three probes in the delayed test.

Note that this was not the case for the high FMC condition, where low probes had significantly lower response time than high probes and medium probes. The result could indicate that the advantage of lower response time for low probes was only present when form-meaning consistency was high and the retrieval of the correct meaning was fast; in contrast, in the medium FMC condition, the retrieval of the correct meaning took longer time because of the lower FMC, so participants may have relied more on their immediate rating for their delayed rating.

Low FMC Against Different Probes

Table 7 juxtaposes the hypotheses and the model estimates for comparing low, medium, and high probes within the medium FMC condition. Figure 11a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 11b shows the model estimated mean ratings for low, medium, and high probes within the medium FMC condition, across the three testing groups, setting within-chimera item order to 1. Figure 11c is the same type of graph for the model estimated mean response times.

Table 7

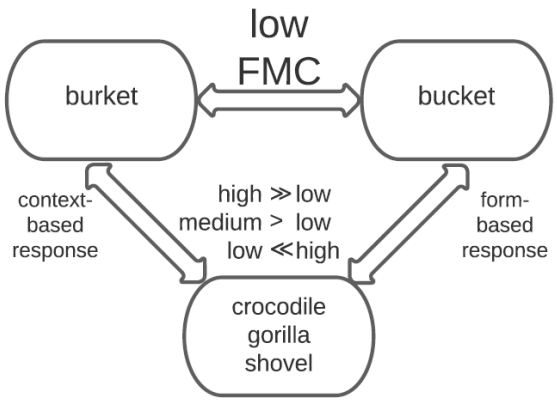
Low FMC Against Different Probes: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
Low FMC	burket-crocodile burket-gorilla burket-shovel	**Ratings:** High probe > Medium / Low Probe	**High Probe** **(1.560, p < 0.001, es = 1.005)** Medium Probe (0.330, p = 0.110, es = 0.213) *Low Probe* *(Reference)*	**High Probe** **(2.664, p < 0.001, es = 1.754)** **Medium Probe** **(0.923, p < 0.001, es = 0.608)** *Low Probe* *(Reference)*	**High Probe** **(2.594, p < 0.001, es = 1.667)** **Medium Probe** **(0.795, p < 0.001, es = 0.511)** *Low Probe* *(Reference)*
		Response Time: Same across probes	Medium Probe (99.566, p = 0.551, es = 0.067) High Probe (66.806, p = 0.688, es = 0.045) *Low Probe* *(Reference)*	**High Probe** **(367.809, p < 0.001, es = 0.598)** **Medium Probe** **(282.367, p < 0.001, es = 0.459)** *Low Probe* *(Reference)*	High Probe (133.985, p = 0.199, es = 0.161) Medium Probe (100.902, p = 0.335, es = 0.122) *Low Probe* *(Reference)*

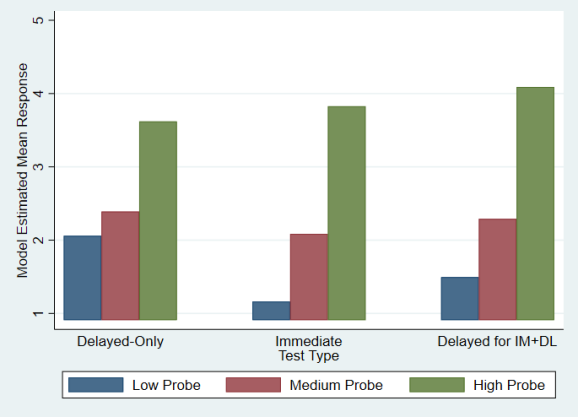
Figure 11

Medium FMC Against Different Probes: Response Discrepancies and Model Estimates

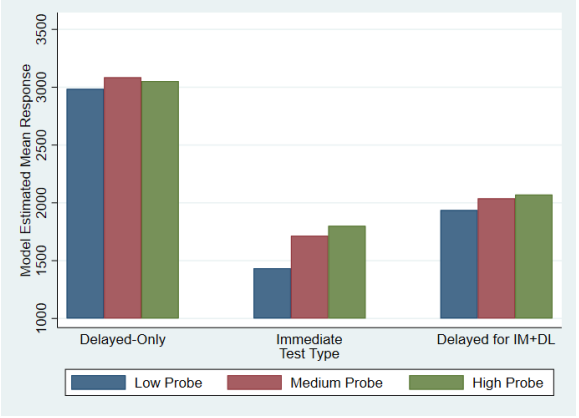
11a



11b



11c



Ratings

Hypothesis. I predicted that for the low FMC condition (e.g., 'burket'), the medium probe (e.g., 'gorilla') would be underestimated because the context-based response (medium) is higher than the form-based response (low; e.g., 'bucket' vs 'gorilla'), whereas the low probe (e.g., 'shovel') would be overestimated because the context-based response ('low') would be much lower than the form-based response (high; e.g., 'bucket' vs. 'shovel'). The degree of the overestimation of the low probe may be higher than the underestimation of the medium probe because the distance between low and high (represented by the double < symbol in Figure 11a) is higher than the distance between medium and low (represented by the single > symbol). The underestimation of medium probes and the overestimation of low probes could decrease the gap between the ratings of these two probes, which may even render the ratings for the medium and low probes comparable to each other.

On the other hand, the ratings of high probes would also be underestimated because the context-based response (high) would be much higher than the form-based response (low; e.g., 'bucket' vs. 'crocodile'). Because both high probes and medium probes would be underestimated, there should still be a gap between high probes and medium probes.

Delayed-Only Group. Model estimates for the delayed-only group confirmed these hypotheses, indicating that the estimated mean rating for high probes was significantly higher than low probes (1.560, $p < 0.001$, es = 1.005) and medium probes (1.230, $p < 0.001$, es = 0.792) whereas medium probe was comparable to low probe (see Figure 11b). Even though the estimated mean rating for medium probes was still numerically higher than low probes, the gap between medium probes and low probes decreased significantly in comparison to all other conditions, as predicted. The medium-low probe difference was estimated to be 0.330 and was statistically insignificant ($p = 0.110$, es = 0.213) for the low FMC condition, in comparison to 0.903 for the high FMC condition ($p < 0.001$, es = 0.581), 1.363 for the medium FMC condition ($p < 0.001$, es = 0.878), and 0.826 for the control condition ($p < 0.001$, es = 0.532) (see Figure 8 under 'Medium FMC Against Different Probes'). Notably, the shrinkage of the medium-low probe difference from high FMC to low FMC was 0.572, which was statistically significant ($p = 0.050$, es = 0.368). The shrinkage of medium-low probe difference from medium FMC to low FMC was 1.032, which was also statistically significant ($p < 0.001$, es = 0.665).

Immediate Tasks for IM+DL Group. For immediate ratings in the low FMC condition, the estimated mean rating of the high probe was also significantly higher than low probes (2.664, $p < 0.001$, es = 1.754) and medium probes (1.741, $p < 0.001$, es = 1.146), as predicted (see Figure 11b). However, there was no significant difference in estimated mean ratings between medium probes and low probes (0.923 points, $p < 0.001$, es = 0.608). This could be because for immediate ratings, it was completely clear to the participants that 'burket' had nothing to do with 'bucket' and was clearly an animal based on the context, so their immediate rating was not influenced by the low FMC at all.

Delayed Tasks for IM+DL Group. For the delayed ratings of the immediate+delayed group, the estimated mean rating of the high probe was also significantly higher than low probes (2.664, $p < 0.001$, es = 1.754) and medium probes (1.799, $p < 0.001$, es = 1.156), as predicted (see Figure 11b). However, similar to the immediate ratings, there was still a significant difference between medium probe and low probe (0.795 points, $p < 0.001$, es = 0.511). This result indicates that rating the relatedness immediately after reading could have mitigated the interference of low

FMC in the delayed ratings because the immediate rating was still fresh in memory. Note that the standardized effect size (0.511) was slightly lower than the immediate ratings (es = 0.608), which indicates that some level of interference from low FMC might be present.

Response Time

Hypothesis. For the low FMC condition, I predicted that the response time would be the same across all probes, because all three probes would have response discrepancy issues as discussed above.

Delayed-Only Group. Model estimates confirmed this hypothesis for the delayed-only group; none of the pairwise comparisons between the three types of probes yielded significant differences (see Figure 11c). This result seems to indicate that the faster response time for low probes observed in the high FMC condition, where no discrepancy existed between the context-based and form-based responses, was offset by the response discrepancy in the low FMC condition (low context-based response vs. high form-based response, see Figure 11a).

Immediate Tasks for IM+DL Group. Similar to the other FMC conditions, in the low FMC condition, the high probes and medium probes were estimated to have significantly higher mean response time than the low probes (see Figure 11c): 1. High Probe > Low Probe (367.809, $p < 0.001$, es = 0.598); 2. Medium Probe > Low Probe (282.367, $p < 0.001$, es = 0.459); 3. High Probe $\approx$ Medium Probe (85.442 ms, $p = 0.247$, es = 0.139). These results indicate that in the immediate ratings, the effect of low FMC was diminished and the generally faster response time on low probes manifested.

Delayed Tasks for IM+DL Group. The delayed test in the immediate+delayed group showed similar results to the delayed test in the delayed-only group. There were no statistically significant differences in estimated mean response times between the three types of probes (see Figure 11c). This finding aligned with that of the medium FMC condition where there were also no differences in response time between the three types of probes. Again, the explanation could be that participants could make equally quick judgments across the three probes in the delayed test because the memory of the immediate ratings was still fresh.

Low Probe Against Different FMC Conditions

Table 8 juxtaposes the hypotheses and the model estimates for comparing low FMC, medium FMC, high FMC, and control conditions within low probes. For the model estimates, low FMC was set as the reference group so that the point estimates of medium FMC, high FMC, and control condition represent the differences in estimated mean ratings or response times between that condition and the low FMC condition, within the low probe.

Figure 12a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 12b shows the model estimated mean ratings low FMC, medium FMC, high FMC, and control conditions within the low probe, across the three testing groups, setting within-chimera item order to 1. Figure 12c is the same type of graph for the model estimated mean response times.

Table 8

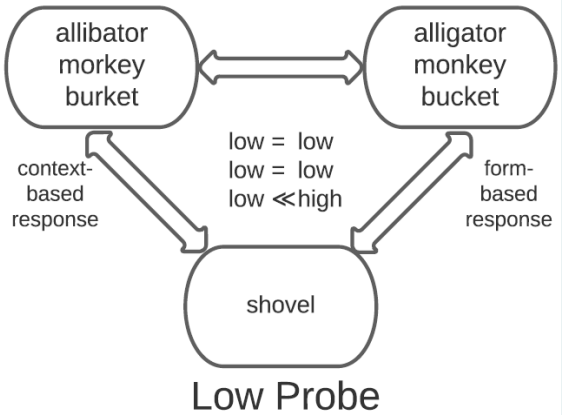
Low Probe Against Different FMC Conditions: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
Low probe	allibator-shovel morkey-shovel burket-shovel darane-shovel	**Ratings:** Low FMC > High / Medium FMC / Control	*Low FMC* *(Reference)* Control (-0.284, p = 0.169, es = 0.183) **High FMC** **(-0.461, p = 0.025, es = -0.297)** **Medium FMC** **(-0.566, p = 0.006, es = -0.365)**	Control (0.005, p = 0.976, es = 0.003) *Low FMC* *(Reference)* Medium FMC (-0.047, p = 0.758, es = -0.031) High FMC (-0.124, p = 0.417, es = -0.082)	*Low FMC* *(Reference)* Control (-0.189, p = 0.246, es = -0.122) Medium FMC (-0.251, p = 0.123, es = -0.162) High FMC (-0.300, p = 0.066, es = -0.193)
		Response Time: Low FMC > Medium FMC > Control > High FMC	*Low FMC* *(Reference)* Control (-192.735, p = 0.282, es = -0.130) **High FMC** **(-614.243, p = 0.001, es = -0.414)** **Medium FMC** **(-655.692, p < 0.001, es = -0.441)**	*Low FMC* *(Reference)* Control (-33.425, p = 0.661, es = -0.054) High FMC (-93.475, p = 0.220, es = -0.152) Medium FMC (-107.914, p = 0.157, es = -0.175)	Medium FMC (1.145, p = 0.991, es = 0.001) *Low FMC* *(Reference)* Control (-74.998, p = 0.479, es = -0.090) **High FMC** **(-340.525, p = 0.001, es = -0.410)**

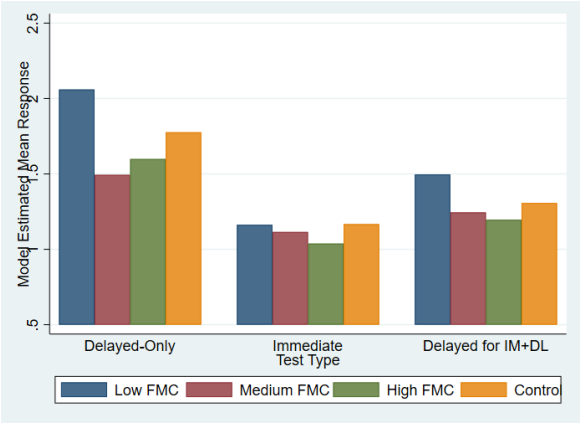
Figure 12

Medium FMC Against Different Probes: Response Discrepancies and Model Estimates

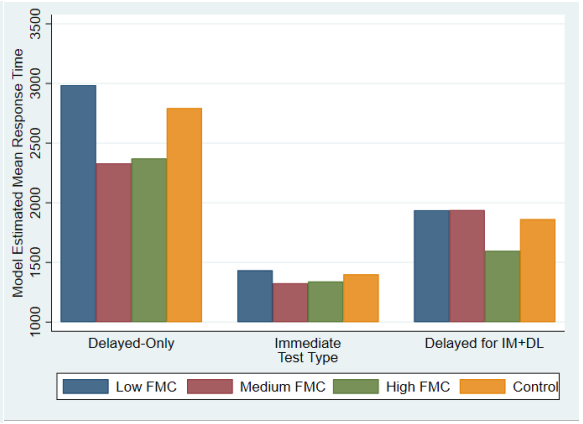
12a



12b



12c



Ratings

Hypothesis. For the low-relatedness probe (e.g., 'shovel'), I predicted that low FMC items (e.g., 'burket' vs. 'shovel') would have an upward bias because the context-based response (low) is lower than the form-based response (high; e.g., 'bucket' vs. 'shovel') (see Figure 12a). I predicted no bias for high FMC (e.g., 'allibator' vs. 'shovel'), medium FMC (e.g., 'morkey' vs. 'shovel'), and control (e.g., 'darane' vs. 'shovel') items because there would be no response discrepancies (see Figures 12a). Therefore, low FMC would lead to higher ratings than high FMC, medium FMC, and control.

Delayed-Only Group. For the delayed-only responses on low probes, the model estimates aligned with the predictions, except for the control condition. The estimated mean ratings for high FMC and medium FMC were significantly lower than low FMC with medium standardized effect sizes (-0.297 and -0.365, respectively).

Interestingly, the estimated mean rating of the control condition was not significantly different from the low FMC condition (-0.284, $p = 0.169$, es = 0.183). The ratings of the control condition may be subject to more randomness than the other conditions because it was a completely novel word form, and participants had the additional cognitive burden of learning the novel word form on top of connecting that word form to meaning. The other conditions, learning the word form was relatively easy because the pseudowords were only one letter different from a familiar word.

Additionally, it is possible that some participants adopted a conservative strategy when rating words in the control condition because they had more difficulty recalling the word's meaning, so they decided to not rate the relatedness too high for the high probe (see discussions below under 'High Probe Against Different FMC Conditions') and not rate the word too low for the low probe. Their responses seemed to be biased towards the neutral rating 3 across low, medium, and high probes (see Figure 8 on Page 32), thus deviating from my prediction that participants' response to the control condition would represent an unbiased 'ground truth.'

Immediate Tasks for IM+DL Group. For the immediate responses, there were no significant differences for any of the pairwise comparisons across the four conditions, and the standardized effect sizes were all very low, as well. This makes sense because when rating immediately, it was very clear that the pseudoword they just encountered had nothing to do with the low probe, thus the upward bias for the low FMC condition was absent.

Delayed Tasks for IM+DL Group. For the immediate+delayed group, the gaps between the low FMC condition and the other three conditions were higher in the delayed ratings than the immediate ratings; however, none of these gaps reached statistical significance, nor were they as high as the gaps in the delayed-only group. The highest gap was between low FMC and high FMC, which was almost significant (-0.300 points, $p = 0.066$) with a small standardized effect size (0.193). This result suggests that the interference effect of low FMC started to manifest in the delayed tasks in the immediate+delayed group even though the memory of the immediate response may have largely mitigated this interference effect.

Response Time

Hypothesis. I predicted that the response time would be the longest for low FMC items (e.g., 'burket' vs. 'shovel'), followed by medium FMC (e.g., 'morkey' vs. 'shovel'), control (e.g., 'darane' vs. 'shovel'), and then high FMC (e.g., 'allibator' vs. 'shovel'). The reason was that a lower FMC pseudoword would activate a real word that is less semantically related to the chimera (e.g., 'burket' activating 'bucket' and 'morkey' activating 'monkey'), thus delaying the retrieval of the correct meaning (e.g., an alligator-like chimera); whereas a high FMC pseudoword would activate a real word that is semantically related to the chimera (e.g., 'allibator' activating 'alligator'), thus facilitating the retrieval of the correct meaning.

Additionally, for low probes, low FMC condition would lead to a discrepancy between the form-based response (low) and the context-based response (high; e.g., 'bucket' vs. 'Shovel'), which would further delay response time; in contrast, the medium and high FMC conditions had no response discrepancy.

Delayed-Only Group. Model estimates support these predictions, except for the control condition. High FMC and medium FMC conditions were estimated to have a mean response time that was 614.243 ms ($p = 0.001$, es = -0.414) and 655.692 ms ($p < 0.001$, es = -0.441) lower than low FMC, respectively. The difference between high FMC and control was also significant (-421.508, $p = 0.018$, es = 0.311)

The difference between high FMC and medium FMC was not statistically significant (41.449 ms, $p = 0.817$, es = 0.027). Neither high nor medium FMC had a response discrepancy issue. This result indicates that medium FMC did not interfere with retrieving the correct meaning as hypothesized. This could be because some participants made a biased inference about the pseudoword's meaning due to the medium FMC and thus, to them, the biased meaning is the correct meaning, thus does not delay meaning retrieval time.

The control condition's response time was comparable to the low FMC condition (-192.735, $p = 0.282$, es = -0.130) potentially due to the same reasons why the control condition's ratings were comparable to the low FMC condition: it takes longer to retrieve the meaning for the control condition because it was an entirely novel word form.

Immediate Tasks for IM+DL Group. For the immediate responses, the estimated mean response time did not differ between the four conditions. The reason could be that one did not need to rely on longer-term memory when rating the relatedness immediately after reading the sentences; regardless of the FMC condition, they were equally quick at deciding that what they had just read had nothing to do with the low probe (e.g., 'burket' had nothing to do with 'bucket').

Delayed Tasks for IM+DL Group. For the delayed tasks in the immediate+delayed group, the response time for high FMC was significantly lower than all other conditions, as predicted: 1. High FMC < Low FMC (-340.525, $p = 0.001$, es = -0.410); 2. High FMC < Medium FMC (-341.669, $p = 0.001$, es = -0.411); 3. High FMC < Control (-265.527, $p = 0.012$, es = -0.320).

Note that for the delayed-only group, the medium FMC's response time was around the same as the high FMC condition; in contrast, the medium FMC's response time was more aligned with the low FMC condition here in the delayed tasks for the immediate+delayed group. This indicates that the interference effect of low FMC was much attenuated because the immediate

ratings were still fresh in the memory. On the other hand, the facilitative effect of high FMC was still present.

Medium Probe Against Different FMC Conditions

Table 9 juxtaposes the hypotheses and the model estimates for comparing low FMC, medium FMC, high FMC, and control conditions within medium probes. Figure 13a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 13b shows the model estimated mean ratings low FMC, medium FMC, high FMC, and control conditions within the medium probe, across the three testing groups, setting within-chimera item order to 1. Figure 13c is the same type of graph for the model estimated mean response times.

Table 9

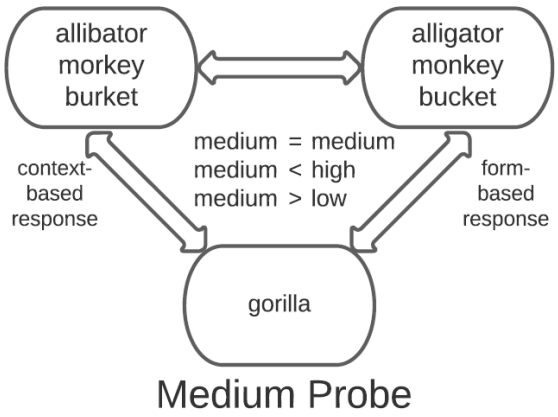
Medium Probe Against Different FMC Conditions: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
Medium probe	allibator-gorilla morkey-gorilla burket-gorilla darane-gorilla	**Ratings:** Medium FMC > High FMC / Control > Low FMC	**Medium FMC** **(0.466, p = 0.024, es = 0.300)** Control (0.212, p = 0.304, es = 0.136) High FMC (0.111, p = 0.589, es = 0.072) *Low FMC* *(Reference)*	**Medium FMC** **(0.350, p = 0.022, es = 0.231)** *Low FMC* *(Reference)* Control (-0.078, p = 0.609, es = -0.052) High FMC (-0.127, p = 0.408, es = -0.083)	**Medium FMC** **(0.418, p = 0.011, es = 0.269)** High FMC (0.226, p = 0.169, es = 0.145) *Low FMC* *(Reference)* Control (-0.266, p = 0.104, es = -0.171)
		Response Time: Low FMC > Medium FMC > Control > High FMC	Control (14.664, p = 0.935, es = 0.010) *Low FMC* *(Reference)* Medium FMC (-207.794, p = 0.245, es = -0.140) High FMC (-301.436, p = 0.092, es = -0.203)	High FMC (24.201, p = 0.751, es = 0.039) Medium FMC (8.992, p = 0.906, es = 0.015) *Low FMC* *(Reference)* Control (-19.811, p = 0.795, es = -0.032)	*Low FMC* *(Reference)* Medium FMC (-36.353, p = 0.733, es = -0.044) Control (-136.553, p = 0.199, es = -0.164) High FMC (-157.204, p = 0.139, es = -0.189)

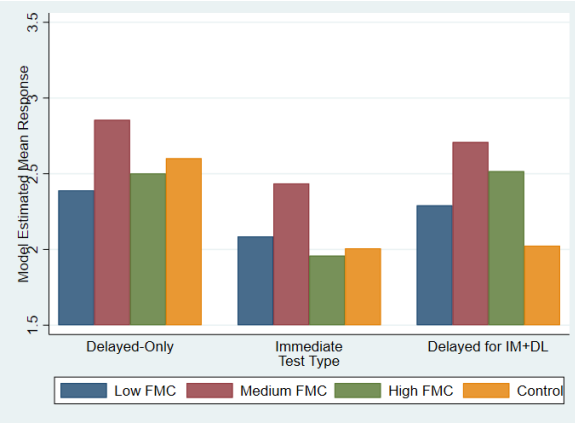
Figure 13

Medium FMC Against Different Probes: Response Discrepancies and Model Estimates

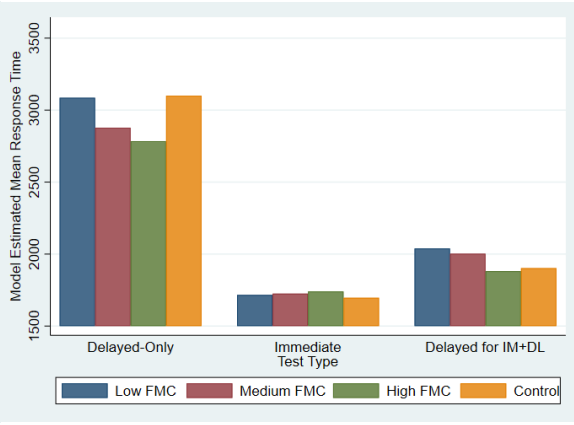
13a



13b



13c



Ratings

Hypothesis. For ratings against medium-relatedness probes (e.g., 'gorilla'), I predicted that medium FMC (e.g., 'morkey') would lead to overestimated relatedness because the context-based response (medium) would be lower than the form-based response (high; e.g., 'monkey' vs. 'gorilla') (see Figure 13a). The high FMC (e.g., 'allibator' vs. 'gorilla') and control (e.g., 'darane' vs. 'gorilla') would be unbiased because of no response discrepancy. The low FMC items (e.g., 'burket' vs. 'gorilla') would be underestimated because the context-based response (medium) would be higher than the form-based response (low; e.g., 'bucket' vs. 'gorilla') (see Figure 13a). Therefore, the ratings would be highest for medium FMC due to the overestimation, followed by high FMC / control (with no bias), and then low FMC due to underestimation.

Delayed-Only Group. Largely aligning with my hypothesis, the model estimated that medium FMC had higher mean ratings than Low FMC (0.466, $p = 0.024$, es = 0.300). The difference between medium FMC and High FMC (0.355 points) was not statistically significant ($p = 0.085$) but had a non-negligible standardized effect size (es = 0.228). This result suggests that the overestimation of medium FMC may require a larger sample with more statistical power to detect.

The estimated mean rating of the Medium FMC was numerically higher (0.254 points) than the control condition, but the difference was also not statistically significant ($p = 0.217$, es = 0.164). As discussed previously (under Low probe against different FMC conditions), this could be because the control condition seemed to be biased towards 3, indicating a neutral rating strategy when the pseudoword was entirely novel and thus word meaning was harder to retrieve.

The estimated mean rating of the high FMC condition was 0.111 points higher than the low FMC condition, but the difference was relatively small, with a standardized effect size of only 0.072, and was not statistically significant ($p = 0.589$). In other words, the downward bias for the low FMC condition seemed to be small in this case, likely because the discrepancy between the context-based response (medium) and the form-based response (low) was relatively small. The 'ground truth' is, on average, around 2 for the medium probes, around 1 for low probes, and around 3.5 – 4 for high probes. The response discrepancy for the medium FMC (i.e., the distance between medium and high, which is around 1.5-2) was higher so the interference effect was more salient; whereas the response discrepancy for the low FMC (i.e., the distance between medium and low, which is around 1) was lower so the interference effect was less salient.

Immediate Tasks for IM+DL Group. The overestimation of medium FMC also showed up in the immediate task (see Figure 13b): 1. Medium FMC > Low FMC (0.350, $p = 0.022$, es = 0.231); 2. Medium FMC > Control (0.429, $p = 0.005$, es = 0.283); 3. Medium FMC > High FMC (0.477, $p = 0.002$, es = 0.314). In other words, even at the stage where the participants were trying to infer the meaning of the pseudoword during reading, they were already biased towards thinking that the pseudoword (e.g., 'morkey') had something to do with the real word activated ('monkey'), even though it did not fit the context well. I intentionally selected the reading material so that the real word activated for the Medium FMC condition would fit the context, but only in an awkward way; the results here indicate that the upward bias still existed despite the relatively poorer fit with the context.

On the other hand, the low FMC was comparable with the high FMC and control conditions. The real word activated ('bucket') from the low FMC pseudoword (e.g., 'burket') clearly did not

fit the context at all so no bias was present at the meaning inference stage, as demonstrated by the immediate response.

Delayed Tasks for IM+DL Group. For the delayed tasks in the immediate+delayed group, medium FMC still had the highest estimated mean ratings, significantly higher than low FMC (0.418, $p = 0.011$, es = 0.269) and control (0.685, $p < 0.001$, es = 0.440). This result indicates that the overestimation caused by medium FMC persisted in the delayed test.

Response Time

Hypothesis. I predicted that for medium probe (e.g., 'gorilla'), low FMC items (e.g., 'burket' vs. 'gorilla') would have the highest response time, followed by medium FMC (e.g., 'morket' vs. 'gorilla'), control (e.g., 'darane' vs. 'gorilla'), and high FMC (e.g., 'allibator' vs. 'gorilla') items. The reasoning is the same as my predictions for the low-probe items.

Delayed-Only Group. For the delayed-only group, the model estimates show a trend aligned with this prediction, except for the 'control' condition (see Figure 13c). Medium FMC and high FMC had lower estimated mean response time than low FMC; even though the differences were not statistically significant ($p = 0.245$ for medium FMC and $p = 0.092$ for high FMC), the standardized effect sizes were not negligible (es = -0.140 for medium FMC and es = -0.203 for high FMC). A more highly powered sample would yield more reliable estimates and determine whether the differences would be significant.

The control condition was estimated to have a response time that was comparable to the low FMC condition ($p = 0.935$, es = 0.010). I found the same result for the low probes, so the reason could also be because recalling the form of an entirely novel word and its meaning took longer time than the other three FMC conditions.

Immediate Tasks for IM+DL Group. Similar to the low probes, for the medium probes, the immediate response time did not differ between the four conditions (see Figure 13c). It was clear to the participants that 'burket' was not related to bucket at all from the reading, thus in the immediate rating, there would be no longer any discrepancy between the context-based response and the form-based response, thus low FMC did not cause any interference on the response time.

Delayed Tasks for IM+DL Group. Similar to their immediate rating, the immediate+delayed group's response time in the delayed tasks also showed no difference between the four conditions (see Figure 13c), suggesting the influence of the immediate ratings on the delayed ratings.

High Probe Against Different FMC Conditions

Table 10 juxtaposes the hypotheses and the model estimates for comparing low FMC, medium FMC, high FMC, and control conditions within high probes. Figure 14a illustrates the hypothesized discrepancies (or lack thereof) between the context-based response and the form-based response for these comparisons. Figure 14b shows the model estimated mean ratings low FMC, medium FMC, high FMC, and control conditions within the high probe, across the three testing groups, setting within-chimera item order to 1. Figure 14c is the same type of graph for the model estimated mean response times.

Table 10

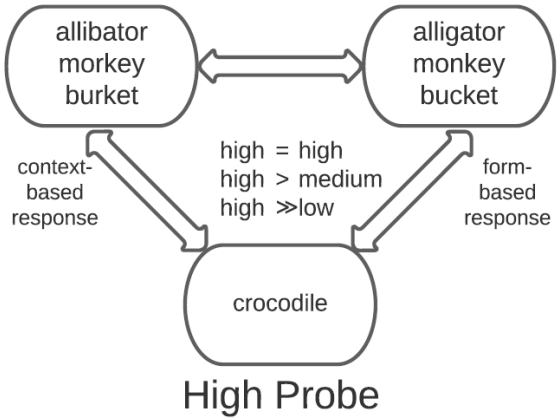
High Probe Against Different FMC Conditions: Comparisons, Hypotheses, and Model Estimates

Types of Comparisons		Hypotheses	Model Estimates		
			Delayed-Only	Immediate for IM+DL	Delayed for IM+DL
High probe	allibator-crocodile morkey-crocodile burket-crocodile darane-crocodile	**Ratings:** High FMC / Control > Medium FMC > Low FMC	High FMC (0.393, p = 0.056, es = 0.253) Medium FMC (0.144, p = 0.485, es = 0.093) *Low FMC* *(Reference)* Control (-0.015, p = 0.941, es = -0.010)	**Control** **(0.308, p = 0.044, es = 0.203)** *Low FMC* *(Reference)* Medium FMC (-0.106, p = 0.490, es = -0.070) High FMC (-0.156, p = 0.307, es = -0.103)	*Low FMC* *(Reference)* Control (-0.014, p = 0.933, es = -0.009) Medium FMC (-0.019, p = 0.908, es = -0.012) High FMC (-0.098, p = 0.548, es = -0.063)
		Response Time: Low FMC > Medium FMC > Control > High FMC	*Low FMC* *(Reference)* Control (-168.748, p = 0.345, es = -0.114) High FMC (-254.288, p = 0.155, es = -0.171) Medium FMC (-309.980, p = 0.083, es = -0.209)	Medium FMC (90.193, p = 0.237, es = 0.147) High FMC (23.253, p = 0.760, es = 0.038) *Low FMC* *(Reference)* Control (-49.781, p = 0.514, es = -0.081)	*Low FMC* *(Reference)* Control (-21.428, p = 0.840, es = -0.026) High FMC (-112.099, p = 0.291, es = -0.135) Medium FMC (-155.373, p = 0.143, es = -0.187)

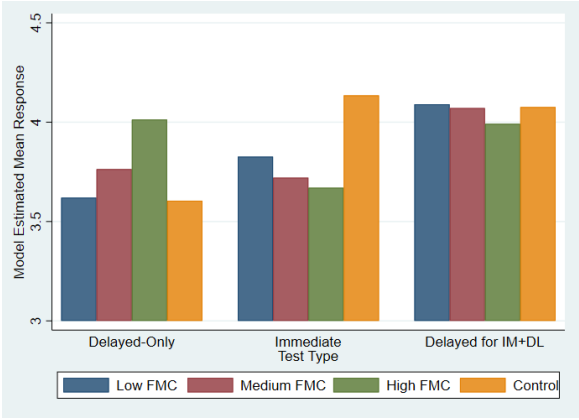
Figure 14

High FMC Against Different Probes: Response Discrepancies and Model Estimates

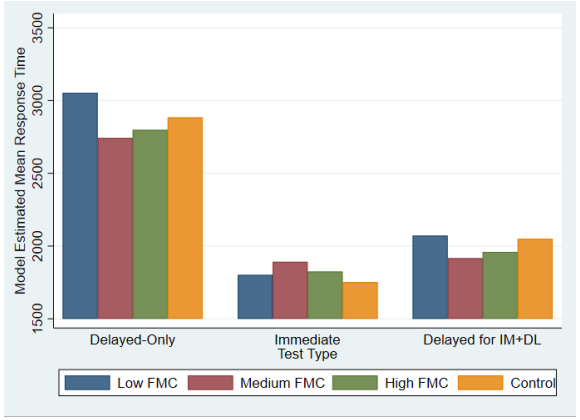
14a



14b



14c



Ratings

Hypothesis. For high probes (e.g., 'crocodile'), I predicted that high FMC (e.g., 'allibator' vs. 'crocodile') and control (e.g., 'darane' vs. 'crocodile') items would yield unbiased ratings, whereas medium FMC items (e.g., 'morkey' vs. 'crocodile') would lead to slight downward bias and low FMC would lead to larger downward bias. This is because, for medium FMC, there would be a slight discrepancy (illustrated as the single > mark in Figure 14a) between the context-based response (high) and the form-based response (medium; e.g., 'monkey' vs. 'crocodile'). In comparison, for low FMC, the discrepancy between the context-based response (high) and the form-based response (low; e.g., 'bucket' vs. 'crocodile') would be larger (illustrated as the double >> mark in Figure 14a), so the bias in rating would also be higher. Therefore, the high FMC and control items would have the unbiased, highest ratings, followed by medium FMC, and then low FMC.

Delayed-Only Group. As shown in Figure 14b, for high probes, the mean ratings were estimated to be the highest for the high FMC condition, followed by medium FMC, and then low FMC, aligning with my predictions. The mean rating of high FMC was estimated to be 0.393 points higher than low FMC, which was marginally significant ($p = 0.056$) with a standardized effect size of 0.253. The mean rating of the medium FMC condition was estimated to be 0.144 points higher than Low FMC, but the difference was not statistically significant ($p = 0.485$) and the standardized effect size was relatively small (es = 0.093). Comparing high FMC and medium FMC, the mean difference in ratings was estimated to be 0.249 ($p = 0.227$, es = 0.160). Even though these differences were not statistically significant based on the current sample, the trend corresponded with my prediction and had non-negligible standardized effect sizes.

The control condition had the lowest rating and was comparable to the low FMC group. This could be owing to the potential neutral rating strategy argued previously (under 'low probe against different FMC conditions' and 'medium probe against different FMC conditions'), where the ratings for the control condition were biased towards 3.

Immediate Tasks for IM+DL Group. Interestingly, the immediate+delayed group's immediate response showed an exactly opposite trend of the delayed-only group, with the control condition being the highest and the high FMC being the lowest. The control condition was significantly higher than all other three FMC conditions, including the high FMC condition. I predicted that medium FMC and low FMC could lead to underestimation due to the response discrepancy, but I did not foresee the underestimation of the high FMC condition. It could be because the participants adopted a conservative strategy because they may have noticed the 'trick' in the varying levels of form meaning consistency and were hesitant about giving a high rating to high FMC-high probe combinations (e.g., 'allibator-crocodile').

Delayed Tasks for IM+DL Group. There were no significant differences between the four conditions in the delayed ratings for the immediate+delayed group. This could be because the high probe provides a very strong cue for recalling the meaning of the pseudoword, which led to equally high ratings across the conditions.

Response Time

Hypothesis. As with the other types of probes, I predicted that the response time would be highest for low FMC, followed by medium FMC, control, and then lowest for high FMC. Additionally, the low FMC condition would lead to the highest discrepancy between the context-based response (high) and the form-based response (low) (represented by the double > symbol in Figure 14a), thus delaying response time the most. In comparison, the discrepancy between the context-based response (high) and the form-based response (medium) was lower for medium FMC (represented by the single > symbol in Figure 14a); thus, the interference would be lower. The high FMC condition had no response discrepancy, and thus the response time would not be affected.

Delayed-Only Group. As predicted, the low FMC condition was estimated to have the highest mean response time (see Figure 14c). Even though the differences between the low FMC condition and the other three conditions were not statistically significant, the standardized effect sizes were not negligible (from -0.114 to -0.209), and thus, a more highly powered sample may help recover the true effects and render more reliable estimates.

The other three conditions (i.e., medium FMC, control, and high FMC) were not significantly different from each other. In other words, medium FMC may not necessarily lead to longer response time than high FMC because of the lower FMC and response discrepancy. Particularly, if participants had made the biased inference that 'morkey' was a monkey-like animal during reading, there would not be any discrepancy between the form-based and context-based responses in their mind, so their response time would not be affected.

Similar to the ratings for low probes and medium probes discussed previously, the control condition's response time being the second highest for rating high probes could, again, be because it took more effort for students to recognize an entirely new word form and recall its meaning, thus leading to longer response time that was close to the low FMC condition.

Immediate Tasks for IM+DL Group. In the immediate task for the high probe, the response time did not differ across the four FMC conditions (see Figure 14c). This result was consistent with the findings for the low probes and medium probes. A similar reason could apply here: Participants did not need to retrieve the word meaning based on the word form in the immediate rating and there would be no response discrepancy involved.

Delayed Tasks for IM+DL Group. The patterns of the delayed tasks in the immediate+delayed group were the same as the delayed-only group, with low FMC having the highest response time, followed by control, high FMC, and medium FMC (see Figure 14c). However, none of these differences were statistically significant. Note that the difference between low FMC and medium FMC was much smaller in the immediate+delayed group's delayed tasks than the delayed-only group and thus were still not statistically significant. In other words, the four conditions seemed to have comparable response time. Again, the reason could be that the immediate rating mitigated the interference of lower FMC in the delayed ratings because the immediate rating was still fresh in memory.

CHAPTER IV: DISCUSSION

In the last chapter, 'Results,' I presented results and a detailed discussion on how these results shed light on the task-specific hypotheses (e.g., 'How did the ratings of low, medium, and high probes differ within the low FMC condition?'). The current chapter 'Discussion' will be a reorganization of the discussions from the last chapter to answer the following broader questions:

1. What are the mechanisms underlying the effect of FMC on task performance? What is the evidence?
2. How is the effect of FMC different for the immediate + delayed group, in comparison to the delayed-only group? What is the evidence?
3. What are the additional, unexpected mechanisms that influenced task performance, in addition to FMC? What is the evidence for the existence of these mechanisms?
4. What are some alternative explanations of the effect of FMC? Do we have evidence for or against these alternative explanations?

Understanding the discussion below requires referring back to the discussion, tables, and figures about task-specific findings in the last chapter. Therefore, at the end of each item of evidence, I will point the readers to a specific section in the last chapter for a more comprehensive understanding of the arguments in the current chapter.

The Effect of FMC on Ratings and Response Time

The findings of this study supported my hypotheses about how FMC would influence the ratings and response time for the semantic learning tasks. Below I will summarize how the empirical findings supported these hypotheses.

Hypothesis 1: Lower degrees of FMC may lead to different types of discrepancy between the context-based and form-based responses, which, in turn, may lead to overestimation or underestimation.

Evidence:

1. Within the medium FMC condition (e.g., 'morkey'), the gap between high probe (e.g., 'crocodile') and medium probe (e.g., 'shovel') decreased in comparison to all other FMC conditions (including control) in all testing groups, indicating the overestimation of the medium probes and underestimation of the high probes, due to the response discrepancy in the medium FMC condition (see 'Medium FMC Against Different Probes').

2. Within the low FMC condition (e.g., 'burket'), the gap between medium probe (e.g., 'gorilla') and low probe (e.g., 'shovel') was not statistically significant and it decreased significantly in comparison to all other FMC conditions (including control) in all testing groups, indicating the underestimation of the medium probes and overestimation of the low probes, due to the response discrepancy in the low FMC condition (see 'Low FMC Against Different Probes').

3. For the delayed-only group's ratings of low probes (e.g., 'shovel'), the estimated mean ratings for high FMC (e.g., 'allibator') and medium FMC (e.g., 'morkey') were significantly lower than low FMC (e.g., 'burket'), indicating overestimation due to the response discrepancy caused by low FMC (see 'Low Probes Against Different FMC Conditions').

4. For the delayed-only group's rating of medium probes (e.g., 'gorilla'), medium FMC (e.g., 'morkey') had higher estimated mean ratings than low FMC (e.g., 'burket') and high FMC (e.g., 'allibator'), indicating overestimation due to the response discrepancy caused by medium FMC (see 'Medium Probe Against Different FMC Conditions').

5. For the delayed-only group's rating of high probes (e.g., 'crocodile'), the mean rating of high FMC (e.g., 'allibator') was estimated to be higher than low FMC (e.g., 'burket') and medium FMC (e.g., 'morkey'), indicating underestimation due to the response discrepancy caused by medium FMC and low FMC (see 'High probes Against Different FMC Conditions').

Hypothesis 2: The higher the response discrepancy is, the larger the bias in ratings may be.

Evidence: For the delayed-only group's rating of medium probes (e.g., 'gorilla'), the difference in ratings between the high FMC condition (e.g., 'allibator') and the low FMC condition (e.g., 'burket') was smaller than the difference in ratings between the medium FMC condition (e.g., 'morkey') and the high FMC condition (e.g., 'allibator'). In other words, for medium probes, the underestimation due to low FMC was smaller than the overestimation due to medium FMC. The reason could be that the response discrepancy (medium > low) caused by low FMC was smaller than the response discrepancy (medium < high) caused by medium FMC, thus leading to smaller bias in ratings (see 'Medium Probes Against Different FMC Conditions').

Hypothesis 3: The response discrepancy caused by lower FMC may lead to longer response time because it takes time to decide which response is correct.

Evidence:

1. Within the medium FMC condition (e.g., 'morkey'), the delayed-only group's response times for the high probes (e.g., 'crocodile') and medium probes (e.g., 'gorilla') were significantly higher than the low probes (e.g., 'shovel'), indicating that the response discrepancy caused by medium FMC for the high probes and medium probes led to longer response time (see 'Medium FMC Against Different Probes').

2. For the delayed-only group's ratings of low probes (e.g., 'shovel'), high FMC (e.g., 'allibator') and medium FMC (e.g., 'morkey') were estimated to have lower mean response times than low FMC (e.g., 'burket'), indicating that the response discrepancy caused by low FMC led to longer response time (see 'Low Probes Against Different FMC Conditions')

Hypothesis 4: The lower the FMC is, the longer the response time may be, above and beyond the effect of response discrepancy because more cognitive resources are needed to suppress the irrelevant semantic information and retrieve the correct meaning.

The findings partially supported this hypothesis, but the evidence was statistically insignificant (but with non-negligible standardized effect sizes) in the current sample. A more highly powered sample is needed for future studies:

1. For the delayed-only group's rating of medium probes (e.g., 'gorilla'), medium FMC (e.g., 'morkey') had slightly lower estimated mean response time (albeit not significant, $p = 0.245$, es = -0.140) than low FMC (e.g., 'burket'). Both low FMC and medium FMC had response discrepancy issues; thus, the faster response time of medium FMC can be attributed to the fact that medium FMC had slightly higher FMC than low FMC (see Medium Probe Against Different FMC Conditions').

2. For the delayed-only group's rating of high probes (e.g., 'crocodile'), medium FMC (e.g., 'morkey') was estimated to have lower mean response time (close to significant; $p = 0.083$, es = -0.209) than low FMC (e.g., 'burket'). Both low FMC and medium FMC had response discrepancy issues; thus, the faster response time of medium FMC can be attributed to the fact that medium FMC had slightly higher FMC than low FMC (see 'High Probe Against Different FMC Conditions').

Mitigated Effect of FMC in the Immediate+Delayed Group

Hypothesis: I predicted that the influence of FMC and response discrepancy on participants' performance would be mitigated in the immediate+delayed group performance. For immediate ratings of low-FMC pseudowords, it should be completely clear to the participants that the pseudoword (e.g., 'burket' as an alligator-rattlesnake chimera) had nothing to do with the real word activated (e.g., 'bucket'), so their immediate rating was not influenced by low FMC at all. Rating the relatedness immediately after reading could have mitigated the effect of FMC in the delayed ratings because the immediate rating was still fresh in memory.

Evidence:

1. For immediate ratings in the low FMC condition (e.g., 'burket'), there was still a significant difference in estimated mean ratings between medium probes (e.g., 'gorilla') and low probes (e.g., 'shovel'), instead of no difference as hypothesized, indicating the hypothesized upward bias of low probes and downward bias of medium probes were not present in the immediate rating. The same pattern persisted in the delayed ratings (see 'Low FMC Against Different Probes').

2. For immediate ratings of low probes (e.g., 'shovel'), the low FMC condition (e.g., 'burket') was comparable with high FMC (e.g., 'allibator'), medium FMC (e.g., 'morkey'), and control (e.g., 'darane') conditions, indicating the hypothesized upward bias was not present in the immediate rating. The same pattern persisted in the delayed ratings. Additionally, in the delayed ratings, medium FMC's response time was close to low FMC. This indicates that the interference effect of low FMC was much attenuated because the immediate ratings were still fresh in the memory (see 'Low Probes Against Different FMC Conditions').

3. For immediate ratings of medium probes (e.g., 'gorilla'), the estimated mean rating of the low FMC condition (e.g., 'burket') was comparable with the high FMC (e.g., 'allibator') and control (e.g., 'darane') conditions, indicating the hypothesized downward bias was not present in the

immediate ratings. Additionally, the estimated mean response time of the low FMC condition did not differ from all other conditions and the same pattern persisted in the delayed ratings (see 'Medium Probes Against Different FMC Conditions').

4. For immediate ratings of high probes (e.g., 'crocodile'), the estimated mean response time of low FMC (e.g., 'burket') did not differ from all other conditions, indicating the hypothesized downward bias was not present in the immediate rating. This pattern persisted in the delayed ratings (see 'High Probes Against Different FMC Conditions').

An interesting finding is that for immediate ratings of high probes (e.g., 'crocodile'), the estimated mean rating of the high FMC condition (e.g., 'allibator') was significantly lower than the control condition (e.g., 'darane'), indicating an underestimation of the high FMC condition. A possible explanation is that participants may have adopted a conservative strategy because they may have noticed the 'trick' in the varying levels of form meaning consistency and were hesitant about giving a high rating to high FMC-high probe combinations (e.g., 'allibator-crocodile').

Additional Mechanisms That Influenced Ratings and Response Times

Not all task-specific predictions aligned with the empirical findings. By analyzing these deviations, I discovered some other mechanisms that affected ratings and response times, in addition to the effect of FMC. Below I will summarize how these additional mechanisms were implicated by the empirical findings.

Mechanism 1: Low-relatedness probes were generally faster to rate than higher-relatedness probes, when none or all of these probes had discrepancies between the context-based and form-based responses.

Evidence:

1. Within the high FMC condition (e.g., 'allibator'), the estimated mean response time of low probes (e.g., 'shovel') was lower than high probes (e.g., 'crocodile') / medium probes (e.g., 'gorilla') across all testing groups (i.e., ratings of the delayed-only group, immediate and delayed ratings of the immediate+delayed group), which deviated from my prediction that the three probes should have equal response times because none of them had response discrepancy issues (see 'High FMC Against Different Probes').

2. Within the low FMC condition (e.g., 'burket'), the estimate mean response time of low probes (e.g., 'shovel') was lower than high (e.g., 'crocodile') / medium probes (e.g., 'gorilla') for the immediate and delayed ratings of the immediate+delayed group, which deviates from my prediction that the three probes should have equal response times because they all had response discrepancy issues (see 'Low FMC Against Different Probes).

Mechanism 2: Recalling the word meaning took more time in the control condition (where the pseudoword is entirely novel and has no real word neighbors) than the other conditions (where the

pseudoword is only one letter different from a familiar word), because it took longer time to recall an entirely novel word form and the meaning associated with that novel form.

Evidence: For the delayed-only group's ratings of all three types of probes (i.e., low, medium, and high probes), the control condition's (e.g., 'darane') response time was comparable to the low FMC condition (e.g., 'burket'), which had the longest response time (see 'Low Probes Against Different FMC Conditions,' 'Medium Probes Against Different FMC Conditions,' and 'High Probes Against Different FMC Conditions').

Mechanism 3: Some participants may have adopted a conservative strategy biased towards the neutral rating (3) for the control condition. The reason could be that they had more difficulty recalling the word's meaning, so they decided against ratings that were too high or too low.

Evidence:

1. For the delayed-only group's ratings of low probes (e.g., 'shovel'), the estimated mean rating of the control condition (e.g., 'darane') was not significantly different from the low FMC condition (e.g., 'burket'), which was overestimated and had the highest rating across all FMC conditions (see 'Low Probes Against Different FMC Conditions').

2. For the delayed-only group's ratings of medium probes (e.g., 'gorilla'), the estimated mean rating of the control condition (e.g., 'darane') was not significantly different from the medium FMC condition (e.g., 'morkey'), which was overestimated and had the highest rating across all FMC conditions (see 'Medium Probes Against Different FMC Conditions').

3. For the delayed-only group's ratings of high probes (e.g., 'crocodile'), the control condition (e.g., 'darane') had the lowest rating and was comparable to the low FMC group (e.g., 'burket'), which was underestimated and the lower than the high FMC and medium FMC conditions (see 'High Probes Against Different FMC Conditions').

Mechanism 4: Some participants may have made a biased inference that the medium-FMC pseudoword (e.g., 'morkey') was related to the real word activated (e.g., 'monkey'), thus for these participants, the activation of the real word did not delay the (biased) meaning retrieval, and the discrepancy between the context-based response and the form-based response was not present.

Evidence:

1. Within the medium FMC condition (e.g., 'morkey'), for the immediate ratings of the immediate+delayed group, the estimated mean response times of the high probes (e.g., 'crocodile') were higher than the medium probes (e.g., 'gorilla'). If the participant had inferred 'morkey' as related to monkey, then the medium probe would no longer have a response discrepancy whereas the high probe would still have a response discrepancy, which explains why the medium probes had lower response time than high probes (see 'Medium FMC Against Different Probes').

2. For the delayed-only group's rating of low probes (e.g., 'shovel'), there was no difference in response time between high FMC (e.g., 'allibator') and medium FMC (e.g., 'morkey'). If they had made the correct inference that 'morkey' was an alligator-like animal instead of a monkey-like one, the medium FMC's response time should be slower than high FMC because the lower FMC should

have interfered with retrieving the correct meaning. However, the two conditions response time ended up being equal, thus indicating that the interference was absent, potentially because they made the biased inference that 'morkey' was a monkey-like animal (see 'Low Probe Against Different FMC Conditions').

3. For the immediate+delayed group's immediate and delayed rating of medium probes (e.g., 'gorilla'), there was an overestimation of medium FMC (e.g., 'morkey') over all other conditions. As discussed previously, the effect of high FMC and low FMC was mitigated in immediate+delayed group's performance. It is, therefore, interesting that the effect of medium FMC was still present, which indicates that the participants made a biased inference of the pseudoword's meaning (see 'Medium Probe Against Different FMC Conditions').

4. For the delayed-only group's ratings of high probes (e.g., 'crocodile'), the estimated mean response time of medium FMC (e.g., 'morkey') was comparable to high FMC (e.g., 'allibator'). This was surprising because medium FMC had response discrepancy whereas high FMC did not; thus, I had predicted that medium FMC would have higher response time. A plausible explanation of the surprising finding is that a biased inference of 'morkey' as a monkey-like animal negated the response discrepancy and the corresponding delay in response time (see 'High Probe Against Different FMC Conditions').

Addressing Alternative Hypothesis

I will address the following alternative explanation for the rating bias caused by low FMC: When rating a low FMC pseudoword against a low probe, the activation of the real word may simply be due to the presence of the highly related probe or at least the co-occurrence of the pseudoword and the probe instead of the pseudoword itself activating the real word. For example, when rating 'burket-shovel', the activation of 'bucket' may simply be due to the presence of 'shovel' or at least the co-occurrence of 'burket' and 'shovel' instead of 'burket' itself activating 'bucket.'

The following findings challenge this alternative hypothesis:

1. For the delayed-only group's ratings of high probes, the mean rating for the low FMC condition (e.g., 'burket-crocodile') was estimated to be 0.393 points lower than the High FMC condition (e.g., 'allibator-crocodile'), which is marginally significant ($p = 0.056$, $es = 0.253$); the mean response time for the low FMC condition was estimated to be 254.288 ms higher than the high FMC condition ($p = 0.155$, $es = 0.171$). While the effect size observed suggests a potential delay in response times for the low FMC condition compared to the high FMC condition, the lack of statistical significance suggests that more data are needed.

2. For the delayed-only group's ratings of medium probes, the mean response time for the low FMC condition (e.g., 'burket-gorilla') was estimated to be 301.436 ms ($p = 0.092$, $es = 0.203$) higher than the high FMC condition (e.g., 'allibator-gorilla').

In these cases, the underestimation or longer response time can only be explained by the activation of 'bucket' from 'burket' per se because neither 'crocodile' nor 'gorilla' should activate 'bucket.' Nonetheless, I did find that the interference effect of low FMC was the highest when rated against low probes (e.g., 'burket' vs. 'shovel'). In other words, the activation of 'bucket'

seemed to be the strongest when rating the 'burket-shovel' relatedness, because of the double activation of 'bucket' from both 'burket' per se and the task (i.e., the co-occurrence of 'burket' and 'shovel'). However, the task-induced stronger activation (and potentially higher interference) did not contradict that there was always a certain level of activation.

Connection with Existing Literature

This study's findings regarding the nuances of FMC in word learning dovetail with the observations of Marelli et al. (2015) and Amenta et al. (2017) regarding the effect of orthographic-semantic consistency (OSC) and phonology-semantics consistency (PSC) on word recognition. Marelli et al.'s finding that high OSC accelerates word recognition resonates with my observation that high FMC facilitates the learning and retrieval of new words by reducing cognitive load and enhancing semantic predictability.

The present study extends the existing literature about the inhibitory effects of semantically unrelated orthographic neighbors on word recognition in semantic tasks among adult native speakers (e.g., Forster & Hector, 2022; Rodd, 2004; Bowers, et al., 2005). My findings demonstrate that the semantic learning of novel words is impacted by the activation of semantically unrelated but orthographically or phonologically similar known words. This interference echoes the challenge participants faced in rejecting 'turple' as a non-animal when its orthographic neighbor 'turtle' was activated (Forster & Hector, 2002). It also aligns with the findings of Rodd (2004), where the semantic activation of orthographic neighbors (e.g., the activation of 'leopard' when seeing 'leotard') caused slower response times in categorizing words as animal or plant names. The extension of these findings to the realm of learning new words underlines the pervasive influence of form-meaning mapping in lexical processing and acquisition.

Moreover, the results echo the difficulty of learning additional, unrelated meanings for homonyms, homophones, and homographs (Casenhiser, 2005; Fang et al., 2017; Mazzocco, 1997; Rodd et al., 2012; Saemen, 1970) and the ease of learning additional, related meanings for polysemous words (Floyd & Goldberg, 2021; Srinivasan et al., 2017, 2019; Srinivasan & Rabagliati, 2021; Srinivasan & Snedeker, 2011). The inhibitory effects of the activation of unrelated semantic representations, as demonstrated in studies on homonymy, find a parallel in my findings on the negative impact of low FMC on word learning. On the other hand, higher FMC facilitates learning, which is parallel to prior studies on the learning of polysemous words. These findings underscore the integral role of form-meaning mappings in efficient linguistic processing and the impact of consistent and inconsistent mappings in language acquisition.

Conclusion

Consistent mapping between form and meaning within language reduces cognitive load for language usage and learning (Dautriche et al., 2017; Kirby et al., 2008; Louwerse & Qu, 2017) and has been found to be a significant predictor of response time in word recognition (Amenta et al., 2017; Marelli et al., 2015; Marelli & Amenta, 2018). I hypothesize that the consistency of the mapping between form and meaning may have implications for word learning as well. English has many word pairs that violate form-meaning consistency and thus may cause learning difficulties, such as homonyms. However, traditional taxonomies of word pairs (such as homonymy, polysemy,

etc.) only consider ‘sameness’ and ‘difference’ but ignore more fine-grained ‘similarity’ or ‘relatedness.’

Therefore, I put forth a new model called Form-Meaning Consistency (FMC) to systematically categorize word pairs according to degrees of similarity between words regarding spelling, sound, and meaning. The FMC model quantifies the degree of form-meaning consistency between word pairs using existing computational metrics of orthographic, phonological, and semantic distances between words.

I designed a novel experimental paradigm to examine the intricate dynamics between FMC and word learning. The evidence garnered through a meticulous examination of semantic relatedness ratings and response times under different FMC conditions and types of probes illuminates the profound influence of FMC on language processing and learning.

The study provided initial evidence for the following hypotheses:

1. High FMC may lead to no bias in rating across probes because of no discrepancy between the context-based and form-based responses.
2. The response time for high FMC may be the shortest among all conditions because retrieval of the meaning may be faster and there is no response discrepancy.
3. Lower degrees of FMC may lead to different types of discrepancy between the context-based and form-based responses, which, in turn, may lead to overestimation or underestimation.
4. The higher the response discrepancy is, the larger the bias in ratings may be.
5. The response discrepancy caused by lower FMC may lead to longer response time because it takes time to decide which response is correct.
6. The lower the FMC is, the longer the response time may be, above and beyond the effect of response discrepancy because more cognitive resources are needed to suppress the irrelevant semantic information and retrieve the correct meaning.

Furthermore, the comparative analysis between the immediate+delayed and delayed-only groups provided insightful revelations about the temporal dynamics of FMC effects. The immediate assessment in the immediate+delayed group seemed to mitigate the FMC effect, suggesting that proximity to word exposure may dampen the interference caused by form-meaning inconsistencies, which persisted, to a degree, in the delayed tasks. This finding underscores the potential for immediate semantic tasks in enhancing word learning efficiency.

This study also discovered additional, unexpected mechanisms influencing task performance beyond FMC. Notably, the phenomenon where low-relatedness probes were generally rated faster across all FMC conditions suggests an underlying cognitive efficiency in dismissing unrelated semantic connections.

Limitations and Future Directions

The sample size of the current study was relatively small (50 in total, including 23 in the delayed-only group and 27 in the immediate+delayed group) and may not be sufficiently powered for some of the statistical comparisons. There were several cases of non-negligible standardized

effect sizes that were not statistically significant. The small sample size increased the risk of Type-II errors. Future larger-scale research will provide more comprehensive and reliable insights into the effect of FMC on the cognitive processes of word learning.

In the current study, the orthographic distance between each pseudoword and its real-word neighbor was 1, and the phonological distance was not considered. Future studies may utilize pseudowords that have different combinations of phonological, orthographic, and semantic distances with a real word. For instance, one can generate pseudowords whose Levenshtein distance from the base real word is 1, 2, or 3, and assign meanings to these pseudowords so that the semantic similarities between the pseudowords and the real words range from 0-1.

Additionally, I created all pseudowords by substituting a letter near the center of a real word. Future research can examine if different positions of the substitution and different methods of creating the pseudowords (e.g., adding or deleting a letter from a real word) may lead to differential effects of FMC on word learning.

Given the finding that the effect of FMC was most salient in the delay-only group, it is plausible to predict that the effect would compound with further delay, indicating that the initial semantic representation formed during reading is impacting the responses, and is slowly decaying and having progressively less of an effect on the delayed response. Future research could examine how the effect of FMC changes at different time points after the initial exposure.

Future research can also explore computational models that may simulate the effect of FMC on human learning. A prime candidate is a sub-lexical distributional semantic model called Fasttext (Bojanowski et al., 2017), which can generate semantic representations for both complete words (lexical units, e.g., 'leo') and partial letter strings (sub-lexical units, e.g., 'leopard'). This model can make inferences on the meaning of pseudowords by combining the semantic representations of partial letter strings, thus representing words with overlapping letter strings as semantically similar.

The current study focuses on how adults acquire novel words through written input. It will be fruitful to explore the effect of FMC on children's acquisition of words through aural input. This will enable us to examine whether the consistency between phonological form (disentangled from orthographic form) and semantics has an effect on learning.

Significance

The FMC model serves as a unified theoretical framework synthesizing diverse lines of research, such as the learning of polysemous words, homonyms, homographs, and homophones. An exciting new line of research is enabled by this model to examine whether and how form-meaning consistency may affect the learning of words with *similar* forms but meanings of various degrees of difference or relatedness.

Additionally, extant studies focus on whether each specific type of word pair may pose learning difficulty instead of comparing the degrees of learning difficulty across different types. For instance, research has shown that homonyms cause difficulty, but is it more difficult to learn homonyms than orthographic neighbors with unrelated meanings? The FMC model treats different types of word pairs as falling under a continuum with different combinations of orthographic,

phonological, and semantic distances, instead of clean-cut categories, which may lead us to a deeper understanding of whether and how form-meaning consistency plays a role in learning.

Methodologically, the innovative experimental design of the present study enables a meticulous examination of the complex interplay between psycholinguistic features and cognitive processes in word learning and memory. Pedagogically, the FMC model and the empirical findings hold promise for identifying words that may be challenging to learn solely from reading, offering guidelines for targeted instruction or scaffolding in educational settings.

This study contributes to a deeper understanding of the cognitive mechanisms underpinning language learning and usage and how they interact with psycholinguistic factors. It underscores the significance of consistent form-meaning mapping not just in word recognition, but crucially in the initial stages of word learning. This extension of the effect of systematicity to the learning domain opens new avenues for exploring how our mind navigates the complex landscape of form and meaning to acquire and process language. By building on the foundational insights of Saussure (1916), Kirby et al. (2008), and Marelli & Amenta (2018), this research enriches the ongoing dialogue about the interplay between linguistic form, meaning, and cognition.

REFERENCES

Amenta, S., Marelli, M., & Sulpizio, S. (2017). From sound to meaning: Phonology-to-Semantics mapping in visual word recognition. *Psychonomic Bulletin & Review*, *24*, 887-893.

Amenta, S., Crepaldi, D., & Marelli, M. (2020). Consistency measures individuate dissociating semantic modulations in priming paradigms: A new look on semantics in the processing of (complex) words. *Quarterly Journal of Experimental Psychology*, *73*(10), 1546-1563.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 1).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"?. *Journal of Memory and Language*, *52*(1), 131-143.

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319-343.

Cassani, G., Chuang, Y. Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 621.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149-2169.

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*, 488-499.

Fang, X., Perfetti, C., & Stafura, J. (2017). Learning new meanings for known words: Biphasic effects of prior knowledge. *Language, Cognition and Neuroscience*, *32*(5), 637-649.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203-12208.

Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 29.

Forster, K. I., & Hector, J. (2002). Cascaded versus noncascaded models of lexical and semantic processing: The turple effect. *Memory & Cognition*, *30*(7), 1106-1117.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Hulme, R. C., Begum, A., Nation, K., & Rodd, J. M. (2023). Diversity of narrative context disrupts the early stage of learning the meanings of novel words. *Psychonomic Bulletin & Review*, 1-13.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349-364.

Kirby, S., Cornish, H., Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677-705.

Longtin, C. M., Segui, J., & Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*(3), 313-334.

Louwerse, M. M., Qu, Z. (2017). Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin & Review*, *24*(3), 849–855.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpusbased methods for extrapolating psycholinguistic variables?. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1623–1642.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*, *50*(4), 1482-1495.

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography–Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, *68*, 1571–1583.

Mazzocco, M. M. (1997). Children's interpretations of homonyms: A developmental study. *Journal of Child Language, 24*(2), 441-467.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Pecher, D. (2001). Perception is a two-way junction: Feedback semantics in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 545–551.

Princeton University. (2010). About WordNet. *WordNet*. Princeton University.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090-1098.

Rodd, J. M. (2004). When do leotards get their spots? Semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin & Review*, *11*(3), 434-439.

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition, 40*(7), 1095-1108.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245-266.

Saemen, R.A. (1970). *Effects of commonly known meanings on determining obscure meanings of multiple-meaning words in context*. Office of Education (DHEW).

Saussure, F. D. (1916). *Course in general linguistics* (Baskin, W., Trans.). Fontana/Collins.

Srinivasan, M., Al-Mughairy, S., Foushee, R., & Barner, D. (2017). Learning language from within: Children use semantic generalizations to infer word meanings. *Cognition*, *159*, 11-24.

Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology: General*, *148*(5), 926.

Srinivasan, M., & Rabagliati, H. (2021). The implications of polysemy for theories of word learning. *Child Development Perspectives*, *15*(3), 148-153.

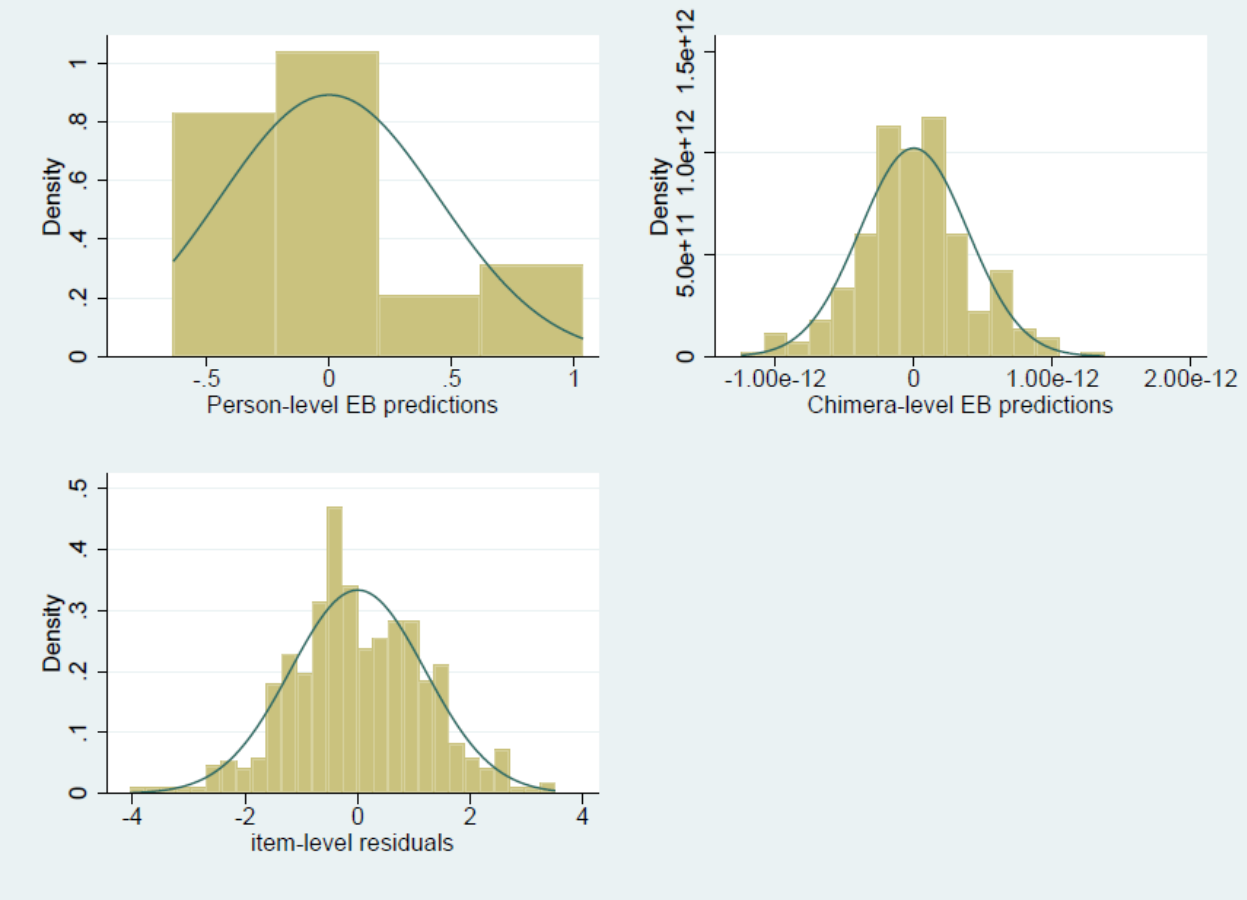
Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, *62*(4), 245-272.

Stella, M., Beckage, N. M., Brede, M., & De Domenico, M. (2018). Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports*, *8*(1), 1-11.

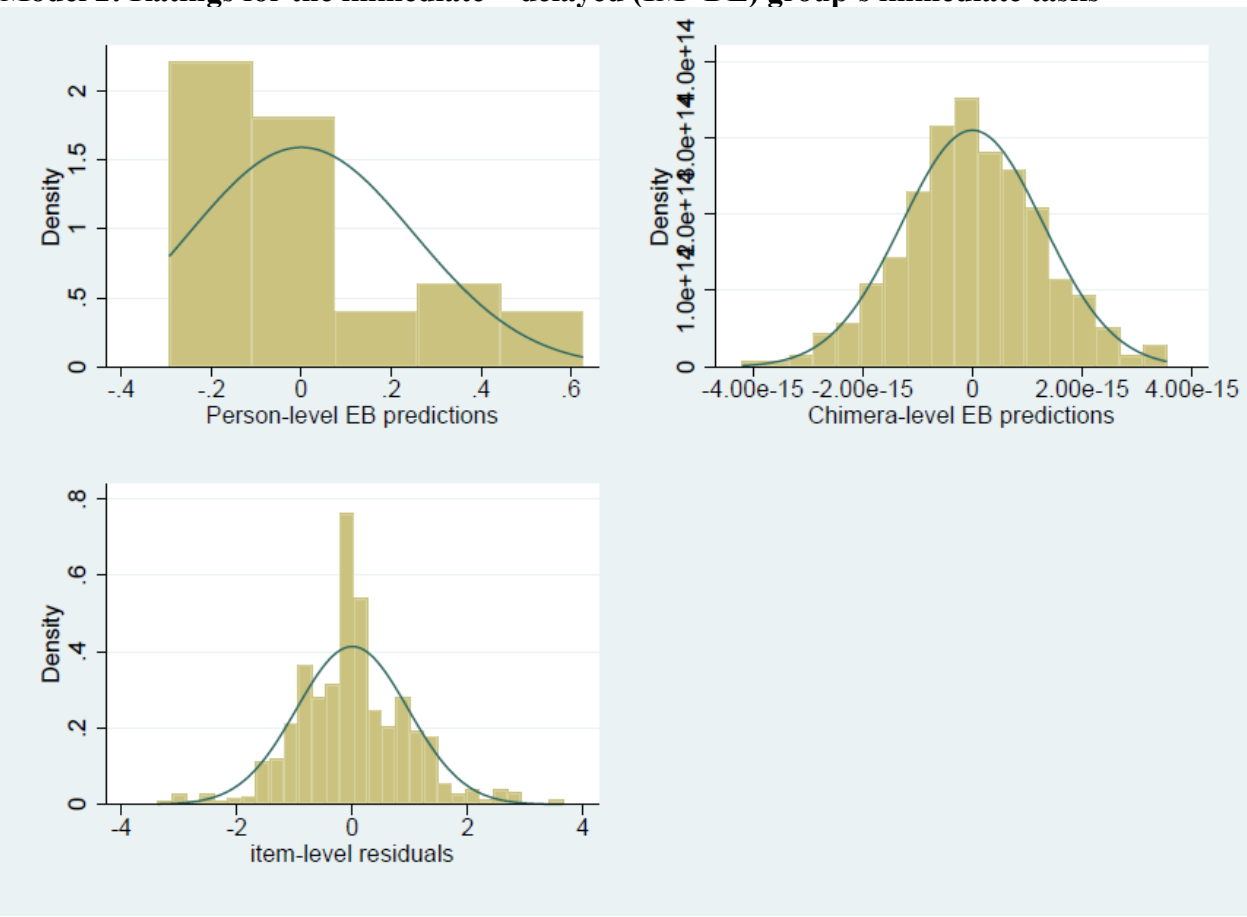
Supplemental Material: Model Diagnostics

Across the six models, the chimera-level random effects (estimated using Empirical Bayes), as well as item-level residuals, all appear to be normally distributed, aligning with the normality assumption. The person-level random effects (estimated using Empirical Bayes) do not follow a strict distribution, due to the limitation of a small sample size (23 participants in the delayed-only group).

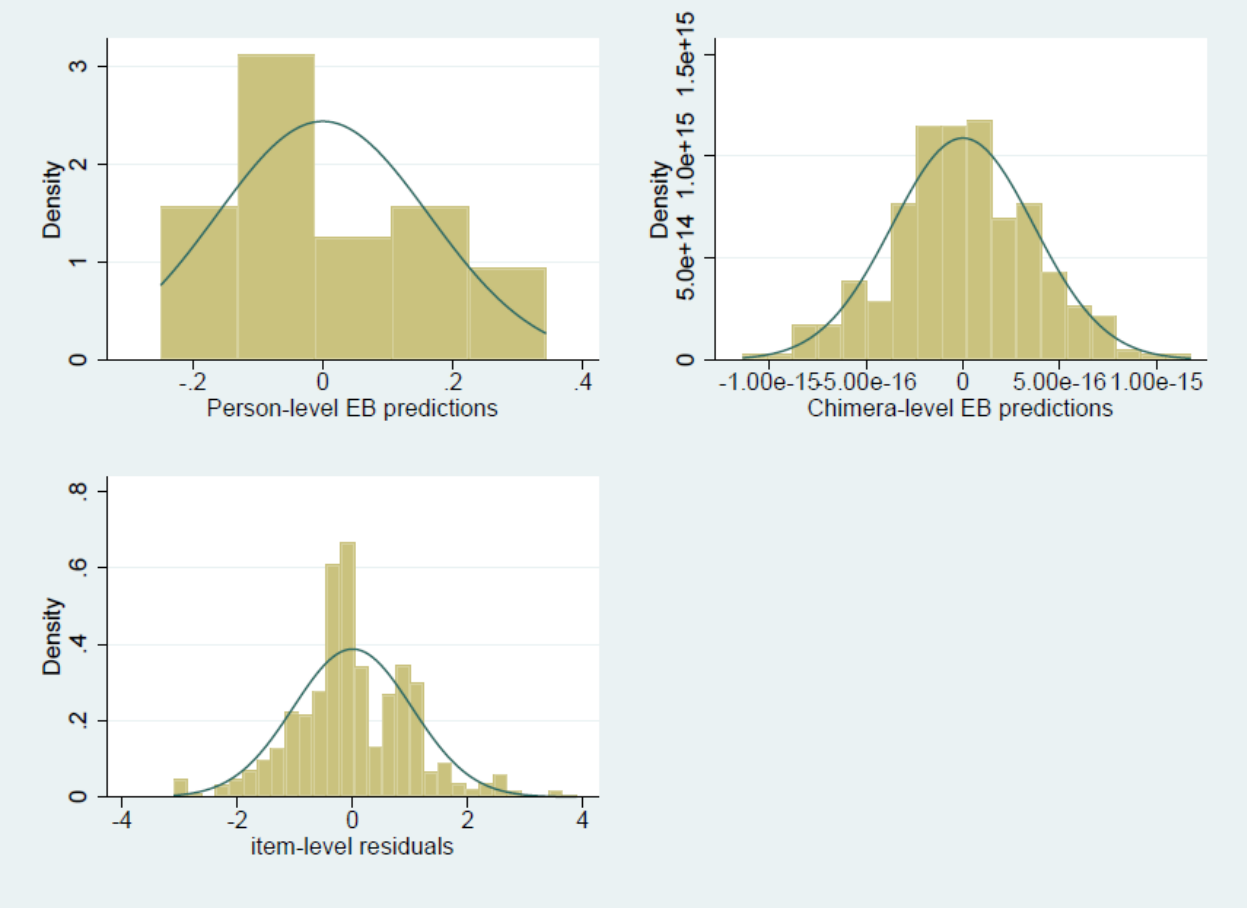
Model 1: Ratings for the delayed-only group



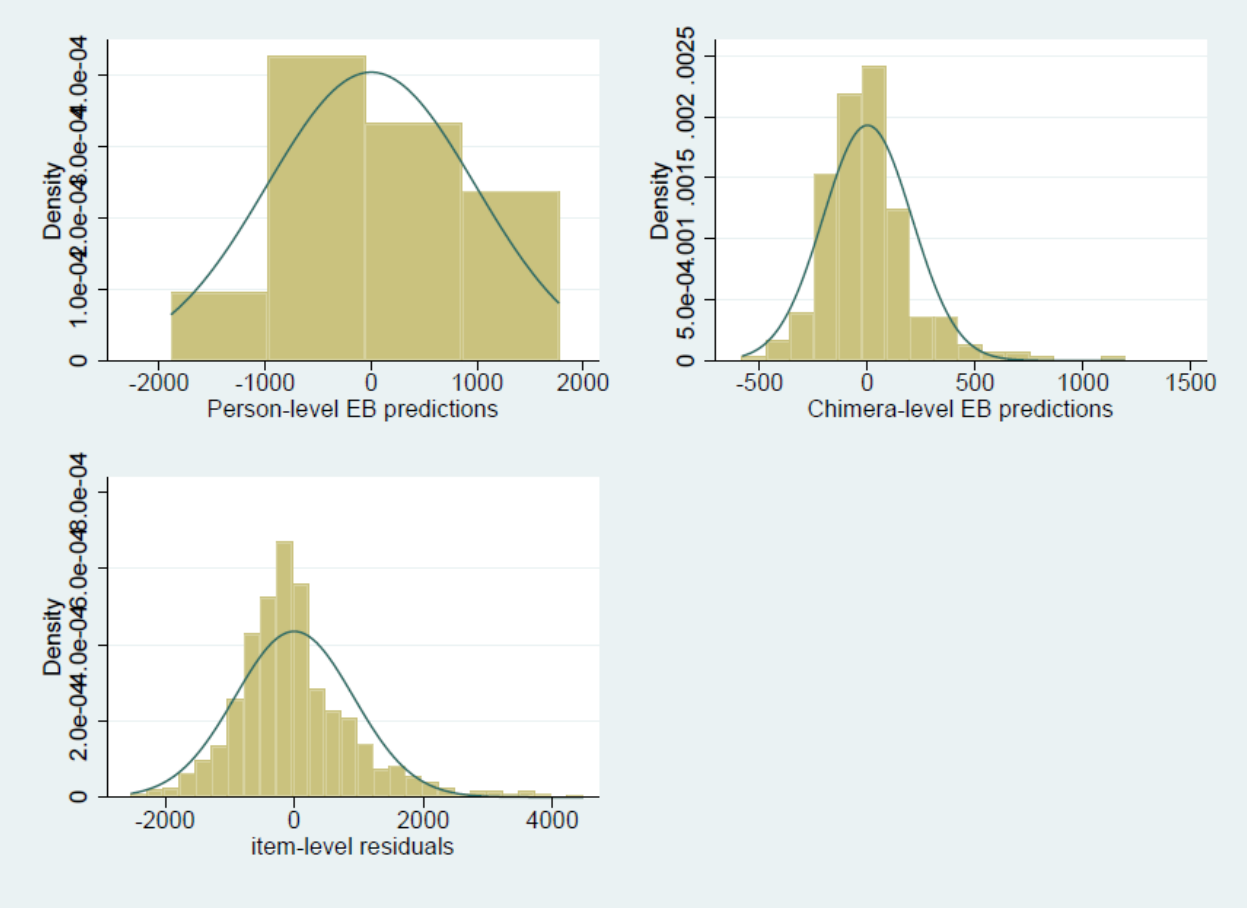
Model 2: Ratings for the immediate + delayed (IM+DL) group's immediate tasks



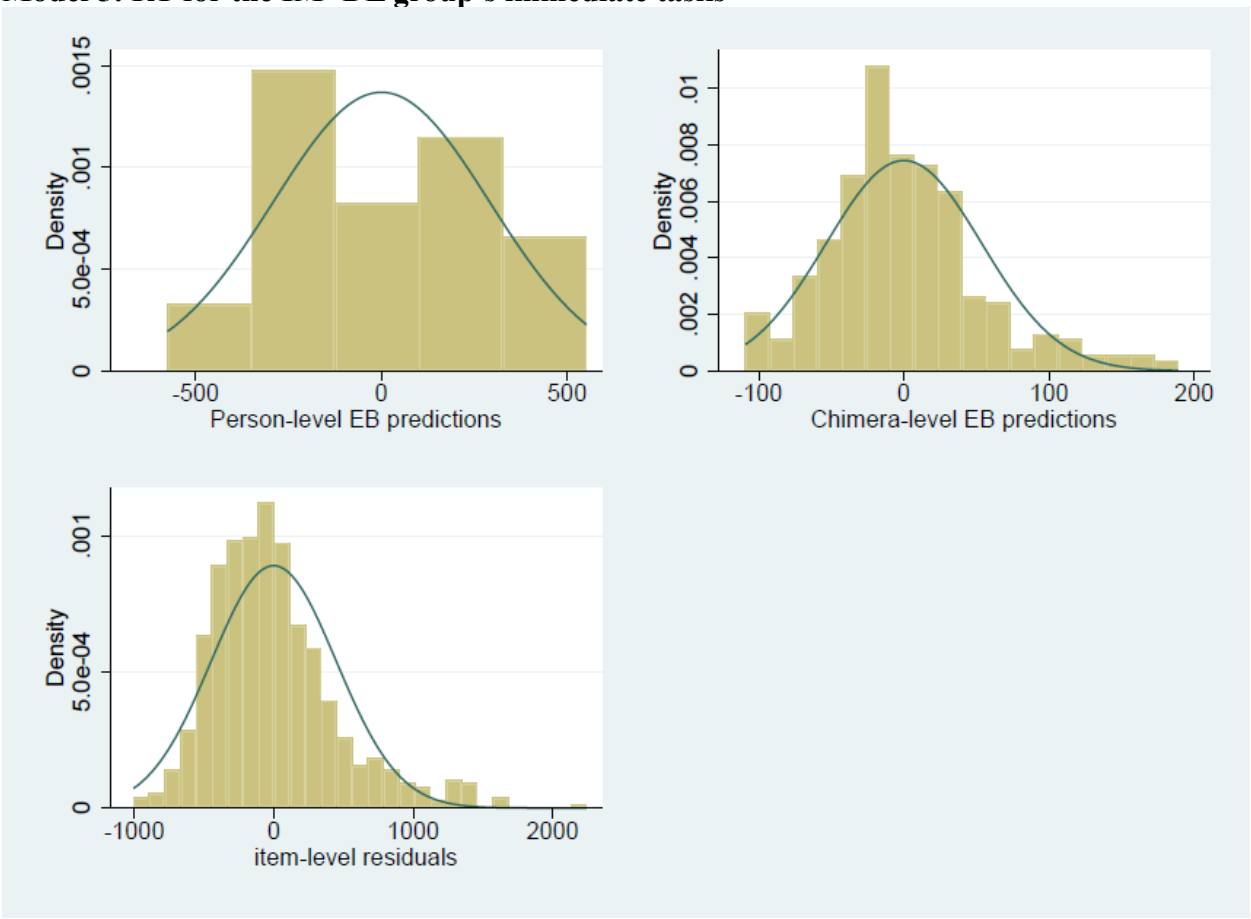
Model 3: Ratings for the IM+DL group's delayed tasks



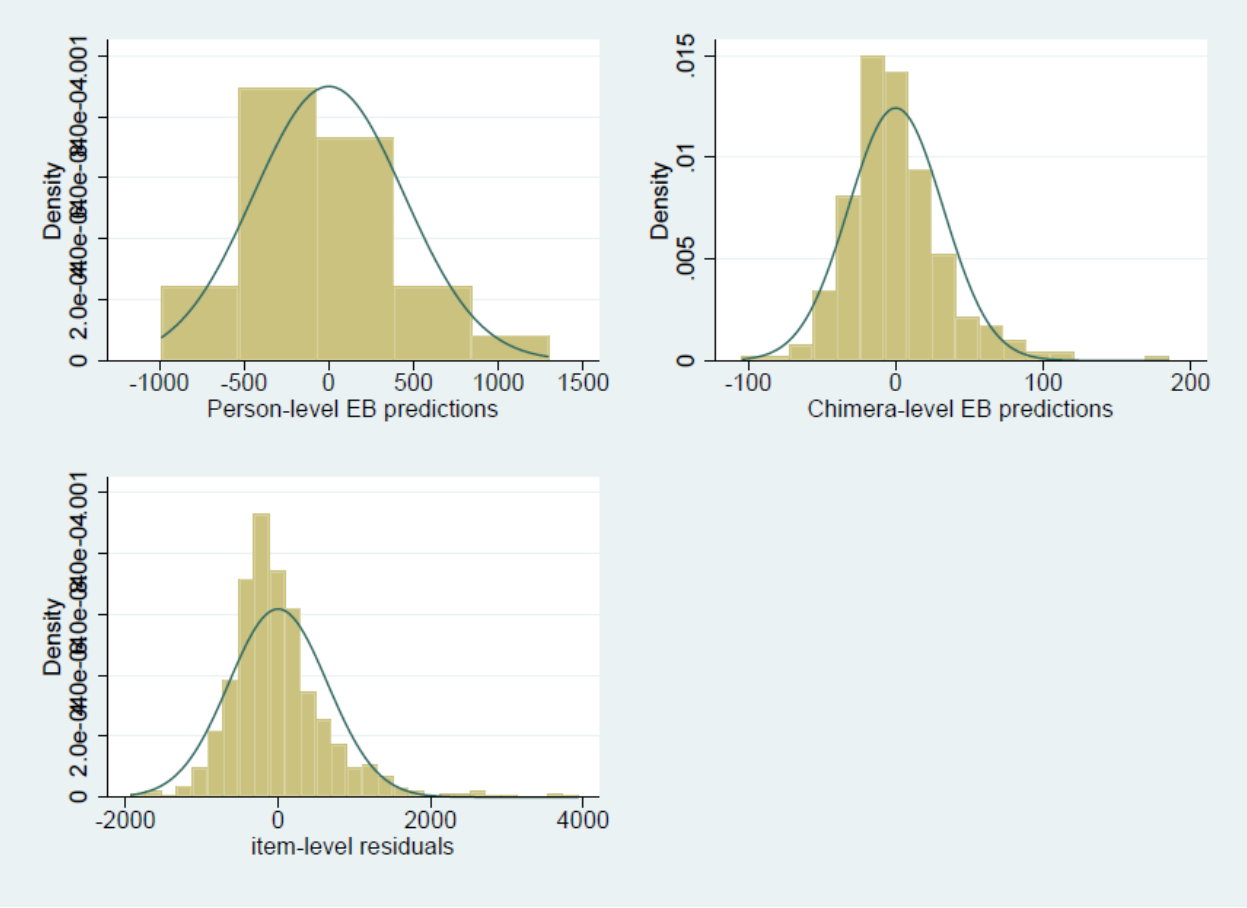
Model 4: Response time (RT) for the delayed-only group



Model 5: RT for the IM+DL group's immediate tasks



Model 6: RT for the IM+DL group's delayed tasks



Supplemental Material: Experimental Stimuli

Following are the probes, base words, and pseudowords for each of the 12 chimeras.

Chimera: potato-turnip

Probe	Base word	Pseudoword
Broccoli	Potato	Porato
Melon	Orange	Oradge
Trout	Salmon	Sasmon
		Vernag (control)

Chimera: cucumber-celery

Probe	Base word	Pseudoword
Onion	Celery	Cekery
Pear	Mango	Mungo
Cushion	Pillow	Pirlow
		Gemack (control)

Chimera: corn-yam

Probe	Base word	Pseudoword
Turnip	Pumpkin	Pumpsin
Peach	Cherry	Chorry
Buffalo	Zebra	Zegra
		Gitid (control)

Chimera: broccoli-spinach

Probe	Base word	Pseudoword
Celery	Spinach	Spirach
Grape	Banana	Balana
Teapot	Bottle	Bontle

		Segost (control)

Chimera: car-van

Probe	Base word	Pseudoword
Jeep	Caravan	Caranon
Skateboard	Bicycle	Bicacle
Parrot	Canary	Capary
		Thrafel (control)

Chimera: train-bus

Probe	Base word	Pseudoword
Taxi	Shuttle	Shultle
Submarine	Ferry	Fefry
Cricket	Scorpion	Scortion
		Nacrut (control)

Chimera: dishwasher-oven

Probe	Base word	Pseudoword
Stove	Furnace	Furtace
Kettle	Barrel	Baurel
Panther	Jaguar	Japuar
		Nefrim (control)

Chimera: cannon-rifle

Probe	Base word	Pseudoword
Bomb	Rifle	Rikle
Spear	Dagger	Dagrer
Lion	Cougar	Coupar
		Rordin (control)

Chimera: alligator-rattlesnake

Probe	Base word	Pseudoword
Crocodile	Alligator	Allibator
Gorilla	Monkey	Morkey
Shovel	Bucket	Burket
		Darane (control)

Chimera: elephant-bison

Probe	Base word	Pseudoword
Reindeer	Elephant	Elethant
Spider	Mosquito	Mostuito
Bolts	Wrench	Wronch
		Naisern (control)

Chimera: peacock-goose

Probe	Base word	Pseudoword
Eagle	Peacock	Pescock
Bear	Cheetah	Cheepah
Cucumber	Garlic	Garnic
		Gleadop (control)

Chimera: caterpillar-cockroach

Probe	Base word	Pseudoword
Beetle	Cockroach	Coctroach
Squirrel	Hamster	Harster
Shack	Cottage	Cothage
		Teissem (control)

Following are the reading materials for each chimera

Chimera: potato-turnip

Reading material (XXX represents the pseudoword; XXXs represents the plural form; depending on the actual pseudoword used, the plural form could end in -s, -es, or -ies)

Melt the butter in a saucepan and cook the shallot and XXXs gently for five minutes, stirring occasionally.

Cut XXXs into anchovy shapes, blanch, oil and salt them and add black peppercorns.

We now have 17 pots of XXXs, all with room to be earthed up at least once.

Grass was conserved as the main stock feed but barley and XXXs were also grown.

A tougher variety known as "hardy XXXs" are generally sown with rape for autumn and winter grazing.

It eliminated the need to grow acres of XXXs or grain to feed the cattle during the long winter months.

In all these areas barley and XXXs are common, but in the east wheat and sugarbeet are grown also.

Radish, XXXs and peas have been sown while the broad beans sown before Christmas are growing away.

Fear of death loomed over the village, for they had not enough XXXs to last until the next harvest.

XXXs are quicker growing than or swedes, but are not frost-resistant and do not keep so well in a clamp.

Chimera: cucumber-celery

Reading material (XXX represents the pseudoword; XXXs represents the plural form)

Arrange the lettuce leaves over a large serving platter with the tomatoes, XXXs, and radishes.

Add the chicken, pasta, remaining XXXs and wine and simmer for 1 minute.

Cut the XXXs into quarters lengthways and scoop out the seeds by running a teaspoon along the center.

Heat the oil over medium heat, add the XXXs and sauté gently for 5 minutes.

For rabbits, use red pimento for ears, strips of XXX peel for whiskers and small pieces of sunflower seeds for eyes.

Deep fry fish cakes and warm for 5 mins before serving with tomato sauce and lightly-boiled leeks and XXXs.

XXXs, radishes and lettuce, are mainly water with the occasional vitamin floating around here and there.

I wasn't allowed to eat anything except lettuce and XXXs and dreadful stuff like that.

XXXs and tomatoes as well as peppers are grown in greenhouses with much higher yields.

Add lots of chopped XXXs, a bay leaf, and peppercorns.

Chimera: corn-yam

Reading material (XXX represents the pseudoword; XXXs represents the plural form)

As well as milling XXX, water-powered mills have been used for weaving and spinning.

They love to eat these XXXs with baked ham which are popular at Thanksgiving.

To serve - place a few knobs of butter over the top of XXXs, sprinkle with parsley and serve immediately.

To kick off we had deep-fried XXX wrapped around a filling of diced pork and Chinese mushrooms.

In 1800 there were no fewer than 8 water-powered mills making woolens and grinding XXX.
These flat coral islands are covered with rich heavy soil well-suited to XXX and taro cultivation.
The percentage of malt, wheat, barley, rye and XXX is also important.
Heat the butter in a heavy-based frying pan until it stops foaming, then add the XXX slices in a single layer.
We 're getting XXX, beef, milk, and flour and dividing into packages and taking into inaccessible places.
After they were harvested, people may grow cassava, groundnuts, and XXXs on the same spot.

Chimera: broccoli-spinach
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
Cook the XXX for 3-5 minutes in boiling, salted water until almost tender.
He also added two side orders - buttered XXX and fries served with bloody Mary sauce.
Sow sprouting XXX, which grows from March to May the following spring, from the middle of the month onwards.
They are herbivores and like lettuce, peas, XXX, and the occasional treat of a chopped prawn.
To add extra flavor to the soup, Steve used the trimmings from the XXX that went into the fish terrine.
Blanch the XXX leaves in boiling water for about 30 seconds, then dry with a paper towel.
All manner of produce fill the fields including yams, shallots, XXX, and asparagus.
When XXX is starting to wilt, add some grated parmesan, stir together and serve soon.
Stir in the XXX, tomatoes, fish, half of the cheese and pepper.
Try tuna, sardines or anchovies, or chopped XXX with plenty of black pepper.

Chimera: car-van
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
The same adult driving a XXX at the same speed might require a thousand times as much power.
By this time the respondent had loaded the goods into a XXX and rejoined Mr. A. in the key department.
A XXX with black-tinted windows appeared at the rendezvous and whisked him inside a palace.
They escaped in the XXX through the same gates they had entered and disappeared into heavy traffic.
At that speed in first gear the engine speed went through the roof and the XXX coasted to a halt.
As they reached the lay-by, the accused had pulled in alongside the red XXX and stopped.
Looking back, I can see one of the porters grappling with the XXX door.
When it was all over we all squeezed into Steve's XXX and drove up to Dingwalls, the club in Camden.
Try to concentrate the weightier items either on, or just ahead of, the XXX’s axle line.
He says that he stopped a XXX, the youths driving it ran off, inside were stolen goods.

Chimera: train-bus
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
He slammed the cabin door, only to hear the hurried application of XXX brakes.
After the cream and brown XXX had rumbled away up the road Constance walked back to the house.

Admire the view from the Ming city walls and explore the bustling Muslim Quarter before boarding the overnight XXX to Lanzhou.
This effort has been strengthened by Police input into driver training schools operated by both major Lothian based XXX companies.
The hum of the traffic was getting louder and every so often the rattling of a XXX set the dirt trembling.
A XXX parked overnight at Templemore School was broken into last Thursday and two fire extinguishers were taken.
On many holidays you may have to carry your own baggage between transfers, on and off XXXs and to your hotel room.
“Shocking Hill!" was the XXX conductor 's cabaret turn on his every approach to Notting Hill.
XXXs are so fast that the passengers complain about not being able to see any of the country’s beautiful scenery.
A gentleman on a bike recently became very agitated when he was "squeezed" out by a XXX.

Chimera: cannon-rifle
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
Against an immobile target, such as a wall, even the early XXX could inflict quite considerable damage.
Humble 'conventional' artillery, XXXs and mortars have killed tens of millions.
The AC-130 Specter gunship flew over the capital several times and blasted arms depots and ammunition stores with 105mm XXX fire and rockets.
Jim goes back inside and loads a XXX, and orders armed guards to be posted around the fence.
I would recommend using the Hellbender more for long range support with the rear XXX and a skilled gunner.
Evidence would be given, he said, that 16 rounds had been fired from one of the XXXs.
But it's different, there's a strange hush in the air and the endless rumbling of 50,000 XXX shells.
The XXX is capable of bringing down a helicopter and has a killing range of more than a mile.
At their head were tanks with XXXs, laced with tear gas, followed by lines of 3,500 riot cops with batons.
Before going to the shooting competition, he visited one of the XXX companies doing selection tests for promotion to Lance Corporal.

Chimera: dishwasher-oven
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
Protein soils such as raw eggs or milk can increase the amount of foam in a XXX.
Place ravioli on dish, cover with sauce, heat in XXX and glaze under grill.
Here are a few suggestions: How many of us know how much water our XXXs use?
This is a top-of-the-range XXX with drop-down door, glass shelf, metal shelf and grid.
Make better use of the space under the sink by fitting a eight-place setting XXX.
In a catering environment the XXX could be used for up to 10 hours a day.
Accommodation: All apartments are air-conditioned and have a kitchenette with 4 ceramic hobs, fridge, and XXX.
Brush with a little egg wash and flash in a hot XXX until golden brown.
Finish off soaked filters by flushing with a hose or pressure jet or passing through a XXX.

It particularly happens in convection or XXXs that circulate dry, hot air around the food in order to cook it faster.

Chimera: alligator-rattlesnake
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
He said Albert reacted like any XXX with live prey, drowning it first and eating it later.
But the kangaroo rat can hear the faint rustles of the XXX's scales moving over the sand, and escape.
Large numbers of XXX skins are exported to Latin America to be made into handbags, shoes and watch straps.
The fangs of this XXX are clearly visible but are not yet in the full striking position.
A widow whose arm was bitten off by a XXX said yesterday she was sorry the creature was later killed.
XXXs eat mice and the young of prairie dogs or cottontail rabbits.
Below it, the greenish water foamed over rocks and there were XXXs lurking in the stony caves along the bank.
Mulder's computer display shows a video of some evil looking hissing XXX from some fact-type website.
The XXX manages to capsize the boat but while Culp disappears beneath the water, Blackmer swims for the surface.
There are a continually galloping rider and an XXX wriggling forwards in the sand that seems to prefigure its destiny.

Chimera: elephant-bison
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
Now Mr Jones fills a mechanical digger with the fruit and lets the XXXs gorge themselves while he continues with his work.
Big gallopers like rhinos have big crests, and so do giraffes and XXXs.
In simplistic terms, the XXX represents strength with the ability to carry a castle on its back.
The great wild bull, the bull of heaven, the wild cow and the XXX bellow.
Meru's XXX population used to be more than 2,000 until the devastating poaching of the 1980's reduced that number to just 300.
Over 25 stalls will offer a huge range of food throughout the festival including XXX steaks, rooster burgers and Leicester curry.
But his pleasure soon turns to distress when he sees that a baby XXX is stuck in the mud and drowning.
Though extinct in the arctic today, endless herds of XXXs were common in the prehistoric northern grasslands of the Pleistocene era.
Sure enough, a huge male XXX was blocking the path ahead, noisily tearing apart a tree.
Stone tools from NW Canada have been found to have traces of XXX blood, by using polymerase chain reaction analysis.

Chimera: peacock-goose
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
A woman with hair the color of the purpliest of XXX feathers was singing on a yard-high rostrum.
Anyway, we fed the ducks and XXXs, Josiah got his hand nipped by an overenthusiastic duck.

Suddenly there was a loud screech, and on the roof a wild XXX appeared.
Then ironically the XXX laid her ten eggs right in the middle of the two trees.
The XXXs don't live long in the wild because of those enormous tails make them easy to catch for predators.
A blue-grey heron glided to rest on a pebbly strand, and a cormorant flew high overhead like a XXX.
A metal disc was fastened between his XXX wings, to serve as a halo behind his head.
A XXX was roasted and carved according to the instructions given.
The gardens within its Moorish walls are populated with walkways, terraces, fountains and XXXs.
At the wildlife sanctuary in Caerlaverock you can see natterjack toads, and in winter flocks of XXXs arrive.

Chimera: caterpillar-cockroach
Reading material (XXX represents the pseudoword; XXXs represents the plural form)
Pied flycatchers breed primarily in broad-leafed woodland which provides the XXXs they require to feed their young.
They found mice droppings in the rice containers and a colony of XXXs in the kitchens.
You can choose from the butterfly, ladybird, bee, green or brown XXX and a slug.
The owner of an Indian take-away was fined $840 after his kitchens were discovered to be infested with XXXs.
The bluetits provided the normal grubs and XXXs, which the blackbird supplemented with juicy worms.
Public health inspectors swooped after a horrified customer spotted a XXX scuttling through the restaurant.
When XXXs are small they can be transferred from the old leaves to the new ones with a paintbrush.
Within 50ms of applying a specific sex pheromone, the InsP-3 level in XXX antennae increases.
Any species of red ant will pick up XXXs, which secrete a sticky sugary substance which ants love.
Maura watched with distaste as Margaret flicked the XXX out of the margarine with the breadknife.